applied sciences

MDPI

Article

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

Maria Cristina Hinojosa Lee *, Johan Braet and Johan Springael

Department of Engineering Management, University of Antwerp, 2000 Antwerpen, Belgium; johan.braet@uantwerpen.be (J.B.); johan.springael@uantwerpen.be (J.S.)
* Correspondence: mariacristina.hinojosalee@uantwerpen.be

Abstract: This study compares various F1-score variants—micro, macro, and weighted—to assess their performance in evaluating text-based emotion classification. Lexicon distillation is employed using the multilabel emotion-annotated datasets XED and GoEmotions. The aim of this paper is to understand when each F1-score variant is better suited for evaluating text-based multilabel emotion classification. Unigram lexicons were derived from the annotated GoEmotions and XED datasets through a binary classification approach. The distilled lexicons were then applied to the GoEmotions and XED annotated datasets to calculate their emotional content, and the results were compared. The findings highlight the behavior of each F1-score variant under different class distributions, emphasizing the importance of appropriate metric selection for reliable model performance evaluation in imbalanced multilabel datasets. Additionally, this study also investigates the effect of the aggregation of negative emotions into broader categories on said F1 metrics. The contribution of this study is to provide insights into how different F1-score variants could improve the reliability of multilabel emotion classifier evaluation, particularly in the context of class imbalance present in the case of phishing emails.

Keywords: emotion analysis; multilabel dataset; annotated datasets; F1-score; performance metrics

check for updates

Citation: Hinojosa Lee, M.C.; Braet, J.; Springael, J. Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores. *Appl. Sci.* **2024**, *14*, 9863. https://doi.org/10.3390/app14219863

Academic Editor: Andrew Teoh Beng Jin

Received: 2 October 2024
Revised: 16 October 2024
Accepted: 24 October 2024
Published: 28 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Machine learning has become an important technology used in diverse sectors, transforming how data are analyzed and utilized [1,2]. Examples of this can be found in fields such as healthcare, financial services, and cybersecurity, where the application of machine learning models has revolutionized the ways risks are predicted, detected, and mitigated [3–5]. The widespread application of both machine learning and deep learning techniques increases the need for new researchers to understand basic statistical principles [6]. This highlights the importance of correctly applying statistical methods and selecting appropriate evaluation metrics to assess model performance [6].

Metrics like precision, recall, and the F1-score originate from disciplines reliant on empirical evidence, including clinical trials, information retrieval, and behavioral research, and have been adapted to evaluate machine learning algorithms [7,8]. In information retrieval, precision and recall measure how well a system retrieves relevant documents requested by the user, whereas in machine learning, relevant documents correspond to positive examples or positive class, and irrelevant documents correspond to negative examples or negative class [9,10].

The consequences of mistakes in evaluating classifiers can be particularly severe in fields such as healthcare, where the costs associated with false negatives (FNs) and false positives (FPs) are significant [11,12]. For instance, a false negative in a disease diagnosis could result in delaying treatment, while a false positive could lead to unnecessary treatments [11]. This underscores the importance of verifying that the metrics used to evaluate machine learning classifiers are appropriate [1]. Poudel explores this and affirms

that diagnosing a disease is a classification task evaluated using performance metrics such as accuracy, precision, recall, and F1-score [13,14]. These four evaluation metrics are commonly used to assess classifiers not only in the medical field but also in areas such as sentiment and emotion analysis. However, researchers often select only one metric, with accuracy being the preferred choice [11,14–16]. Nevertheless, given the cost sensitivity of FN and FP, it is important to consider metrics like precision and recall, which provide a more nuanced evaluation of a classifier's performance.

The presence of class imbalance in training datasets can result in the majority group being overclassified while examples from the minority class are more frequently misclassified [17,18]. When classes are imbalanced, accuracy may be misleading, presenting high scores that do not represent reliable performance [14,17]. Therefore, when working with an imbalanced dataset, it is recommended to use multiple performance metrics [14,18]. Still, criticism persists regarding the use of recall, precision, and the F1-score in text classification, as they overlook the accurate classification of true-negative examples and fail to account for performance that could result from chance [7,19,20].

Sentiment and emotion analyses frequently encounter the challenge of class imbalance, representing a key obstacle to improving sentiment and emotion classifiers [21,22]. The challenge of class imbalance in emotion analysis datasets can lead to models that perform well on majority classes but poorly on minority ones, as some emotions are significantly more represented than others [21,23,24]. This issue is evident with the use of accuracy as an evaluation metric when there is class imbalance, as classifiers tend to give more importance to the majority class, complicating the classifier's ability to perform effectively on the minority class [23].

Metrics like accuracy and the micro and macro F1-score are recommended for evaluating the performance of multilabel classifiers [6]. The recent literature shows that accuracy and the F1-score are the most used performance metrics for multilabel classifiers [6,25]. After reviewing evaluation metrics for binary classifiers, particularly in relation to class imbalance, we identified a lack of research comparing different F1-score variants (micro, macro and weighted), in the context of multilabel emotion analysis [11,26,27]. This gap is important because multilabel emotion analysis often involves imbalanced datasets, and addressing it can help in selecting the right evaluation metric to assess classification models' performance effectively.

The contribution of this study is to provide insights into how different F1-score variants could improve the reliability of multilabel emotion classifier evaluation, particularly in the context of class imbalance present in the case of phishing emails. In this paper, the authors consider that machine learning encompasses both traditional and deep learning models, such as BERT. This research serves as the starting point of a comprehensive research project aimed at improving phishing email detection by using emotion analysis. This is also expected to contribute to more nuanced detection models that incorporate emotional understanding to identify phishing attempts. Part of the context of this research project can be found in the related work by Perkisas [28].

This paper begins with the Section 2, in which an analysis of the relevant literature on performance metrics and emotion classification is provided. The Section 3 details the experiment setup, and the specific performance metrics compared. The Section 4 provide a comprehensive comparison of F1-score variants.

2. Related Work

Sentiment analysis and emotion analysis are subfields of natural language processing (NLP). Sentiment analysis focuses on evaluating whether data carry positive, negative, or neutral sentiments, while emotion analysis is more nuanced and seeks to identify specific emotions within data [29,30].

Emotion theories are generally categorized into three main types: basic emotion theory (BET), continuous models, and componential models [31,32]. Ekman's model is an example of BET, which proposes that emotions are discrete and independent [32]. However,

researchers acknowledge that these basic emotions can interact with each other and with higher-order cognitive processes, resulting in complex emotional experiences beyond simple, discrete categories [33].

Emotion-labeled datasets contain data labeled with corresponding emotion categories [34]. While some researchers create custom emotion taxonomies, Ekman's model remains one of the most commonly used frameworks for labeling emotion datasets [29,34]. This study used the GoEmotions and XED datasets, both available in the Hugging Face repository [35]. Both datasets were chosen for their broad vocabulary and non-social-media content, thus reducing potential biases associated with Twitter-based datasets, which may not generalize well across contexts that are less specific such as emails [36].

The GoEmotions multilabel dataset was published in 2020 and consists of 58,000 manually emotion annotated English comments from Reddit [37]. It is a multilabel dataset that includes 27 emotional labels plus a neutral category [37]. The authors of the paper curated the messages to reduce profanity and harmful content [37]. The annotation process involved three raters, with additional raters included in cases of disagreement. If annotators were uncertain about the emotion in the message, they could select the "neutral" label. The 27 emotions were hierarchically grouped into Ekman's six basic emotions by the authors: anger, disgust, fear, joy, sadness, and surprise [37].

In the case of the XED dataset, it includes 25,000 Finnish and 30,000 English emotion-annotated movie subtitles. The annotation framework uses Plutchik's Wheel of Emotions as a base, which includes anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. University students conducted the annotations, requiring a consensus from at least two of three annotators for each labeled movie subtitle [36].

Evaluation Metrics

Sensitivity or recall (Equation (1)) is the fraction of actual positive cases that are correctly predicted as positive or pertaining to the class. It is also known as the true-positive rate [38–40]. Precision (Equation (2)) is the fraction of correctly predicted positive cases from all cases predicted as positive [38–40]. The F1-score, also known as the Dice similarity coefficient, is the harmonic mean of precision and recall, providing a balance between the two [23,38,39]. For this paper, the micro (Equation (3)), macro (Equation (4)), and weighted (Equation (5)) variants were calculated to compare their performance on the same classification task.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

where TP is the true positive, FP is the false positive, and FN represents the false negative.

Micro F1-score: This metric considers all classes by using the total true positives, false negatives, and false positives:

$$F1_{\text{micro}} = \frac{2 \times Precision_{\text{micro}} \times Recall_{\text{micro}}}{Precision_{\text{micro}} + Recall_{\text{micro}}} \tag{3}$$

Macro F1-score: This metric calculates the F1-score independently for each label and then averages them, giving each class the same weight:

$$F1_{\text{macro}} = \frac{1}{N} \sum_{i=1}^{N} \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{4}$$

Weighted F1-score: This metric calculates the F1-score for each class as in the macro F1-score, but the average is weighted by the number of true instances each class has:

$$F1_{\text{weighted}} = \sum_{i=1}^{N} w_i \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{5}$$

where w_i is the proportion of true instances for label i. The weighted F1-score considers the support or weight of each class, w_i (Equation (6)):

$$w_i = \frac{TP_i + FN_i}{\sum_{j=1}^{N}(TP_j + FN_j)} \quad (6)$$

where TP_i and FN_i are the true positives and the false negatives for class i, and the denominator is the sum of all the true positives and false negatives for all classes.

These evaluation metrics were selected because they are commonly used in assessing the performance of emotion analysis models (see Table 1), utilizing emotion-annotated datasets ranging from news headlines to book passages, movie subtitles, and messages from Twitter and Reddit [36,41–44]. It is important to note that some researchers still rely solely on accuracy [41,43] or the micro F1-score [42]. Using these measures to evaluate performance can result in overly optimistic outcomes or complicate comparisons across different models and datasets [36,45,46].

Table 1. Metrics used in emotion-annotated datasets. The symbol μ indicates that the micro F1-score was used, instead of the macro F1-score and N/A indicates that the information is not available (Öhman et al., 2020) [36].

Study	Source	Number of Emotions	Multi	Macro F1	Accuracy	Is It Balanced?
Abdul-Mageed and Ungar [43]	Twitter	8	No	N/A	95.68%	No
Abdul-Mageed and Ungar [43]	Twitter	24	No	87.47%	N/A	No
Liu et al. [44]	Books, etc.	8 + neu	No	60.4% μ	N/A	Yes Only after dropping disgust and surprise.
Demszky et al. [37]	Reddit	27	Yes	46%	N/A	No
Demszky et al. [37]	Reddit	6	No	64%	N/A	No
XED [36]	Movie Subtitles	8 + neu	Yes	53.6%	54.4%	Yes

In this paper, we consider several variants of the F1-score as well as precision–recall curves. The goals of this paper are to improve the evaluation of text emotion classification models by applying different performance metrics and to compare the effectiveness of micro, macro, and weighted F1 metrics in reflecting the performance of emotion classification models. The objective of the study is to demonstrate the comparative usefulness of the different F1-score variants in the evaluation of text emotion classifiers for imbalanced datasets. The outcomes of our analysis are presented in the results section.

3. Methodology

The following subsections detail the methodology employed.

3.1. Data Preprocessing

Regarding the GoEmotions dataset, we used the Ekman level of emotion classification proposed by the authors of the original paper, excluding the neutral category. This resulted in a dataset of 38,237 messages [37]. The neutral class was not considered for this paper because, as noted in the source paper, it did not include messages with no emotion or with all emotions present, but did include messages that were difficult to classify as containing a specific emotion.

For the XED dataset, the emotions considered were those that are part of the Ekman emotion theory, as this was the classification also used in the GoEmotions dataset: anger, disgust, fear, joy, sadness, and surprise [36,47]. The emotion labels 'anticipation', 'trust', and 'neutral' were excluded.

For this dataset, the granularity of the annotations is at the sentence level; however, the shortest movie subtitle present is "!" and the longest subtitle includes three sentences [36].

Due to this discrepancy, we considered the level of granularity to be at the message level. From this dataset, only the English messages were considered, which resulted in the use of 13,682 English messages (see Table 2).

Table 2. Emotion distribution of messages used from each dataset. This represents the quantity of messages per emotion that were used after removing the 'neutral' class in both datasets, and the 'anticipation' and 'trust' classes in the XED dataset.

Datasets	Origin	Number of Messages	Emotions					
			Anger	Disgust	Fear	Joy	Sadness	Surprise
GoEmotions	Reddit forum	38,237	7021	1012	929	21,730	4030	6668
XED	Subtitles from movies	13,682	3828	2317	2439	2832	2464	2442

A frequent problem with data is that they can be skewed, with the majority of examples pertaining to one class [23,48,49]. When this occurs, the dataset is imbalanced [48,49]. A balanced dataset has a similar number of cases in each class [49]. It is important to determine whether a dataset is balanced because predictive models developed using imbalanced datasets tend to perform better for the largest class [49].

Table 2 shows that, except for the emotion label 'anger', the XED dataset presents more balanced classes compared to the GoEmotions dataset. The GoEmotions dataset exhibits class imbalance; the number of messages labeled as 'joy' in GoEmotions is more than twenty times greater than that of message labeled as 'fear'. This imbalance may stem from using Ekman's emotion classification, where 'joy' is the only positive emotion included [37]. Considering this, we created two datasets: GoEmotions Aggregated and XED Aggregated. These datasets consolidate the emotions into three categories: 'surprise', 'joy', and 'negative emotions'. The 'negative emotions' category combines the emotions 'anger', 'disgust', 'fear', and 'sadness'. Table 3 provides a detailed description of these datasets. The objective of aggregating the negative emotions into one category is to see if there are changes in the F1-scores when classes are more balanced or when negative emotions are part of the same label.

Table 3. Emotion distribution of the messages used from both datasets after aggregating all messages that contain negative emotions into the 'negative' category.

Datasets	Origin	Number of Messages	Emotions		
			Negative	Joy	Surprise
GoEmotions Aggregated	Reddit forum	38,237	12,942	21,730	6668
XED Aggregated	Subtitles from movies	13,682	9425	2832	2442

Both datasets underwent identical preprocessing to standardize the text data for consistent analysis. Using the SpaCy library [50], stopwords and non-alphanumeric characters were removed from the messages, and the text was lowercased. The idea behind removing stopwords is that they are frequently used words that might not convey information and may reduce accuracy and performance if they are not eliminated [51].

3.2. Lexicon Distillation Process

This process extracts and categorizes unique tokens from the preprocessed messages. Instead of considering the frequency of each token, lexicon distillation classifies tokens based on their association with specific emotions. The process involves several necessary steps:

Input: Dataset

1. Tokenization: Tokenize each message from the dataset into individual words.
2. Preprocessing: Remove tokens that are stopwords or non-alphanumeric and lowercase the text.
3. Calculate the token-emotion association and Threshold Setting: Choose a threshold value in [0, 1]. A threshold of 0.0 indicates that a token must appear in at least one message labeled with a particular emotion to be assigned that emotion in the lexicon. Any presence of the token in messages associated with an emotion is sufficient for labeling the token with it.
4. Token Analysis: Determine each token's association with a particular emotion based on the percentage of messages containing the token labeled with that emotion. Identify the messages containing the token. If the percentage of these messages meeting or exceeding the threshold is labeled with a particular emotion, assign that emotion to the token in the lexicon. Repeat this analysis for each token across all emotions. Include tokens that do not meet the threshold for any emotion in the lexicon with a zero value for each emotion.
5. Assign the token to emotion or discard it.
6. Lexicon compilation: Compile the evaluated tokens into a lexicon. Represent each token with a six-component vector, where each component corresponds to one of the six emotions. Association with the emotion is given in a Boolean manner with 0 meaning no association, and 1 meaning association.

Output: Distilled Lexicon

Figure 1 shows a flow chart that describes the process of distilling a lexicon from an annotated dataset.

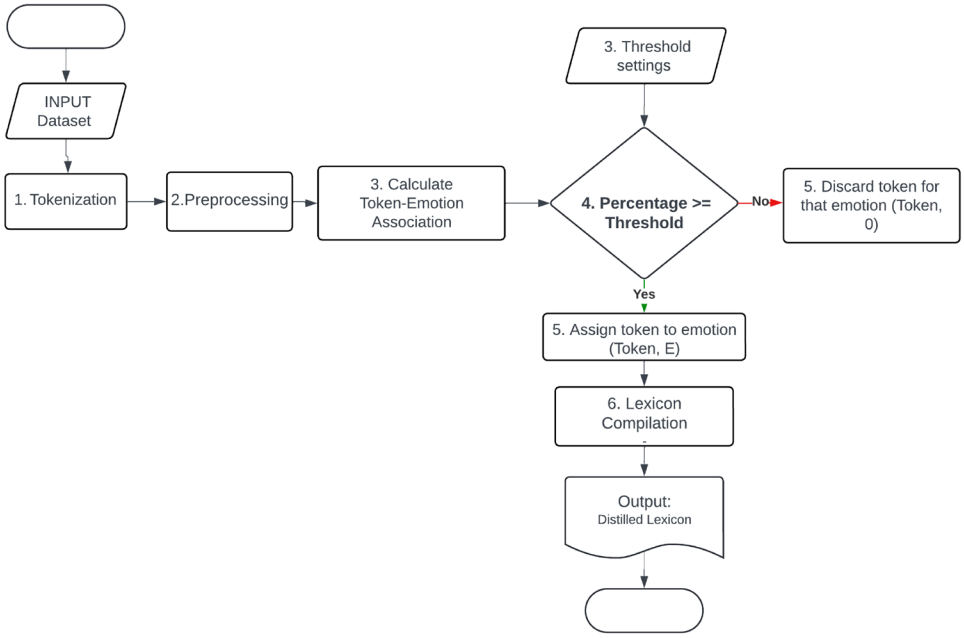


Figure 1. Flow chart describing the process of distilling a lexicon from an annotated dataset.

3.3. Application of Lexicons

Each compiled lexicon was applied to both the GoEmotions and XED datasets, as well as their aggregated versions, to facilitate the evaluation of different F1-score variants: micro, macro, and weighted. The weighted F1-score is calculated by determining the F1-score for each class, defined as the harmonic mean of precision and recall, and then multiplying each result by a weight that considers the number of true instances or true positives (TPs) for each class, and then averaging the outcomes [52]. The weighted F1-score is used when there are imbalanced classes and shows a bias toward the majority class [7,52]. Conversely,

the macro F1-score treats all classes equally by assigning the same weight to each class, meaning that the minority class is as influential as the majority class [52]. Consequently, the model is penalized if it performs poorly on the minority class [52,53]. The micro F1-score is calculated over the entire dataset; this means that it disregards class membership [54].

The process of applying the lexicon into a dataset is as follows:

Input data: The distilled lexicon and the dataset.

1. Tokenization and Stopword Removal: Stopwords are removed, and each message from the dataset is tokenized.
2. Token Evaluation: Each message is scanned to find tokens present in the lexicon.
3. Labeling Emotions: If a token is not in the lexicon, the next token is evaluated. Each message is labeled with the emotions associated with the tokens present in both the text and the lexicon. If tokens are linked to different emotions, or if a single token is linked to multiple emotions, the message reflects all detected emotions. Labeling is achieved through a vector, where each component corresponds to one of the emotions (six for the full model, three for the aggregated model).
4. Matrix Update: The vectors are filled with binary values (0 or 1), indicating the absence or presence of each emotion
5. Comparison with Original Labels: After all tokens have been analyzed, the resulting NxM matrix is compared with the original annotated message labels.

Output: the performance metrics, the graphics and, the dataset.

This process is visually summarized in Figure 2.

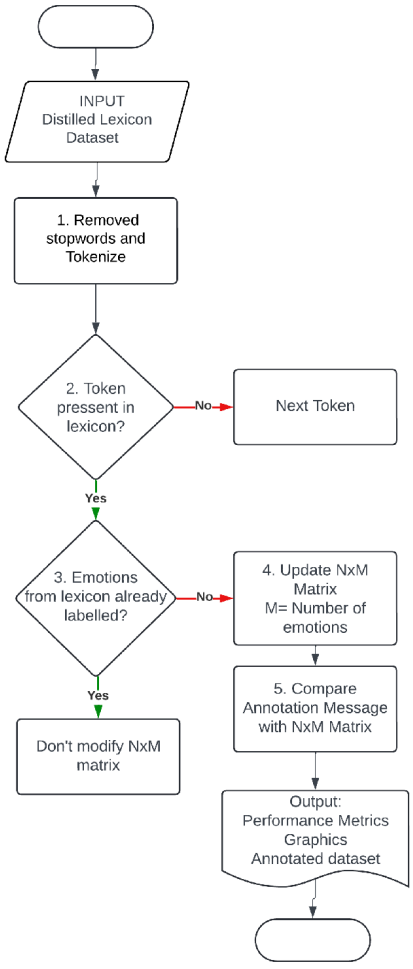


Figure 2. Flow chart describing the process of applying a distilled lexicon to an annotated dataset.

Subsequently, confusion matrices were calculated for each emotion and threshold. A confusion matrix is a tool used to evaluate the performance of a classification task [7,11,55]. It is divided into four segments:

1. Number or percentage of examples that are correctly classified as part of a class (true positives or TPs);
2. Number or percentage of examples that are correctly classified as not being part of a class (true negatives or TNs);
3. Number or percentage of examples that are incorrectly classified as being part of a class (false positives or FPs);
4. Number or percentage of examples that are incorrectly classified as not being part of a class (false negatives or FNs).

Calculating the confusion matrices assesses the model's classification performance. At the same time, the values that form these matrices are the basis for calculating the F1 measures under varying conditions [38,56]. The structure of a confusion matrix is shown in Figure 3.

Actual label	Emotion present	True Positives (TP)	False Negatives (FN)
	No emotion present	False Positives (FP)	True Negatives (TN)
		Emotion Present	No emotion present
	Predicted label		

Figure 3. Confusion matrix structure (Sun et al., 2009) [55].

The comparison of the micro, macro, and weighted F1 metrics is conducted by analyzing the results of the distinct types of F1-scores found after applying the distilled lexicon achieved at various thresholds to emotion-annotated datasets. The specific metrics considered in this paper are recall, precision, and the micro, macro, and weighted F1-scores. These performance metrics are obtained from the data present in the confusion matrix [12].

The F1-score variants are then calculated with the Equations (3)–(6). Acceptable F1-score values vary depending on the domain, task, and specific dataset characteristics. For instance, in fraud detection for mobile applications, a good F1-score is greater than 0.7 [57]. In medical contexts, benchmark studies on annotated colon cancer pathology reports have reported F1-scores ranging from 0.82 to 0.93 for primary tumors and 0.65 for metastatic tumors [58].

In emotion analysis, studies have achieved macro F1-scores of 0.53, 0.62, and 0.64 [36,37,59] using techniques such as BERT, which employs transformer-based models pretrained and fine-tuned to achieve state-of-the-art results [60]. Comparing against a baseline model helps determine what constitutes an acceptable F1-score [61]. For this experiment, we will consider the F1-score—micro, macro, or weighted—of 0.53 achieved in the original paper from the XED dataset as our benchmark for a good F1-score. It is important to note that while BERT considers contextual information, the lexicons used in our experiment only consist of unigrams [36,60].

Precision–recall (PR) curves were also generated for each emotion category, reflecting the varying thresholds at which the lexicons were generated. Figure 4 illustrates the theoretical behavior of precision–recall curves under different scenarios. It could serve as a general reference for understanding curve shapes. The optimal curve would be one that is closer to the top-right corner, indicating a higher area under the curve and better performance in terms of precision and recall.

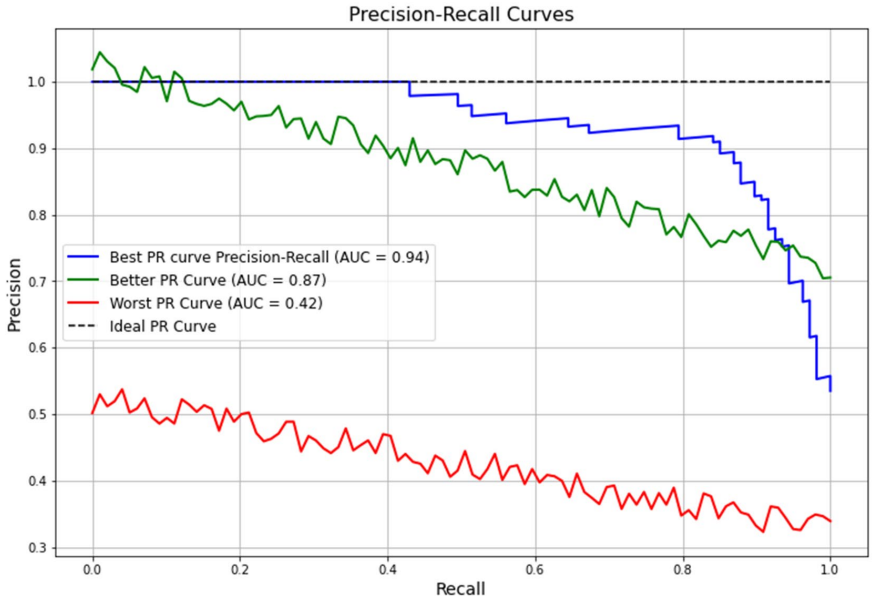


Figure 4. Examples of precision–recall (PR) curves.

4. Results and Discussion

This section presents the resulting micro, macro, and weighted F1-scores of emotion classification tasks on the content of the XED and GoEmotions datasets and their aggregated versions when applying distinct lexicons obtained using different thresholds in [0, 1]. The objective is to determine which F1-score variant best captures classifier performance within these datasets under varied conditions, such as emotion categories and dataset class imbalance.

Each distilled lexicon was applied to the datasets to classify the messages per emotion. Afterwards, the F1-scores were calculated for each emotion category, with thresholds specific to each lexicon. The confusion matrices are available as Supplementary Materials, providing further details on model performance.

4.1. Overview of Experiments

In this study, several experiments were conducted with the GoEmotions and XED datasets and their aggregated versions, to evaluate the performance of different F1-score variants—micro, macro, and weighted—using lexicon-based emotion analysis models. The key experiments involved applying derived lexicons at different thresholds, to various datasets. The experiments are summarized in Figure 5.

Lexicon Source	Dataset Applied	Aggregation Level	Purpose
GoEmotions	GoEmotions	Non Aggregated	Assess performance on same dataset
GoEmotions	XED	Non Aggregated	Evaluate generalizability across datasets
XED	XED	Non Aggregated	Assess performance on same dataset
XED	GoEmotions	Non Aggregated	Evaluate applicability on another dataset
GoEmotions Aggregated	GoEmotions Aggregated	Aggregated	Evaluate the effect of emotion aggregation
GoEmotions Aggregated	XED Aggregated	Aggregated	Assess generalizability of aggregated data
XED Aggregated	XED Aggregated	Aggregated	Assess performance on aggregated dataset
XED Aggregated	GoEmotions Aggregated	Aggregated	Evaluate applicability on another aggregated dataset

Figure 5. Matrix summarizing the experiments conducted.

4.2. Precision–Recall Curves

Precision–recall (PR) curves were generated for each of the six emotions from Ekman and for the three aggregated emotion categories. The points on each curve represent different thresholds used in the evaluation.

Figure 6 displays the precision–recall curves of all six emotions of the GoEmotions distilled lexicon, applied to the GoEmotions dataset. In the case of the emotion "joy", depicted in Figure 6d, the curve demonstrates high recall and precision, indicating better model performance for this emotion. Similarly, Figure 7 illustrates the precision–recall curve for GoEmotions Aggregated distilled lexicon and applied to the GoEmotions Aggregated dataset. It shows that when negative emotions are combined, the model's performance improves, evidenced by the curve's proximity to the top-left corner, signaling high precision and recall levels.

Figure 8 depicts the precision–recall curves for the GoEmotions distilled lexicon applied to the XED dataset. This visualization reveals notably low recall and precision across the curves, indicating that the model is not effective for this particular application.

Conversely, Figure 9 illustrates the precision–recall curve for the GoEmotions Aggregated distilled lexicon applied to the XED Aggregated dataset. In this scenario, the curve demonstrates an improvement in model performance, as evidenced by higher recall values and precision for the aggregated negative emotions (see Figure 9c).

Figure 10 displays the precision–recall curves for the XED distilled lexicon applied to the XED dataset. The curves demonstrate high recall and precision for 'joy' (see Figure 10d) and the other emotions, indicating stronger model performance across different emotional categories. The precision–recall curves for the XED Aggregated distilled lexicon and applied to the XED Aggregated dataset (Figure 11) also show an improvement for the negative emotions, as shown in Figure 11c, compared to the non-aggregated curves.

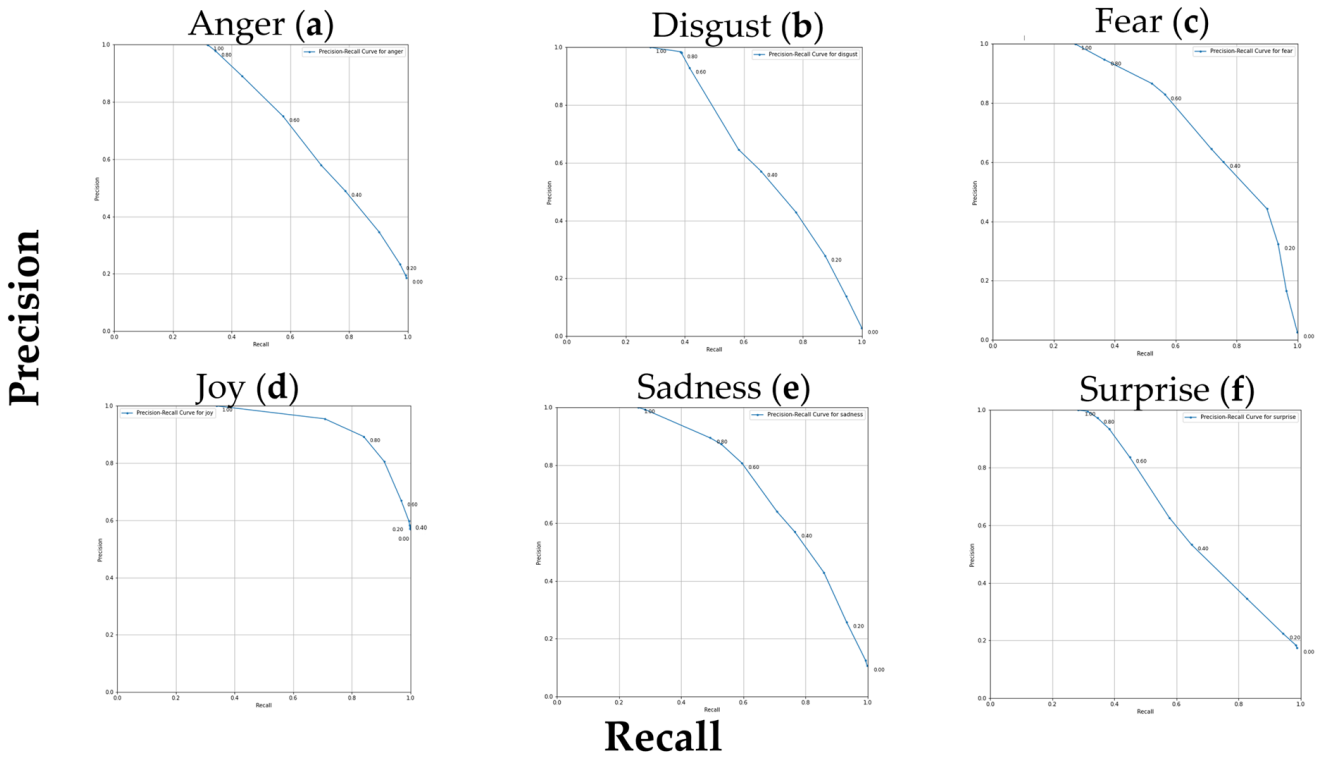


Figure 6. Precision–recall curves for six different emotions analyzed using the GoEmotions distilled lexicon and applied to the GoEmotions dataset. Each graphic represents the precision–recall curve for a specific emotion, with points on the curve representing different thresholds. Emotions include (**a**) anger, (**b**) disgust, (**c**) fear, (**d**) joy, (**e**) sadness, and (**f**) surprise.

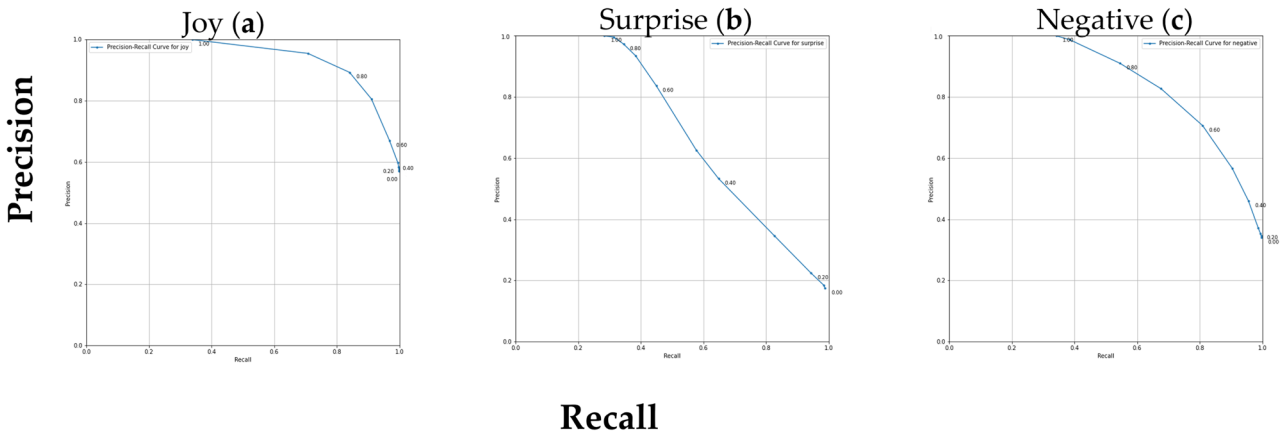


Figure 7. Precision–recall curves for three different emotions analyzed using the GoEmotions Aggregated distilled lexicon and applied to the GoEmotions Aggregated dataset. Each graphic represents the precision–recall curve for a specific emotion, with points on the curve representing different thresholds. Emotions include (**a**) joy, (**b**) surprise, (**c**) negative.

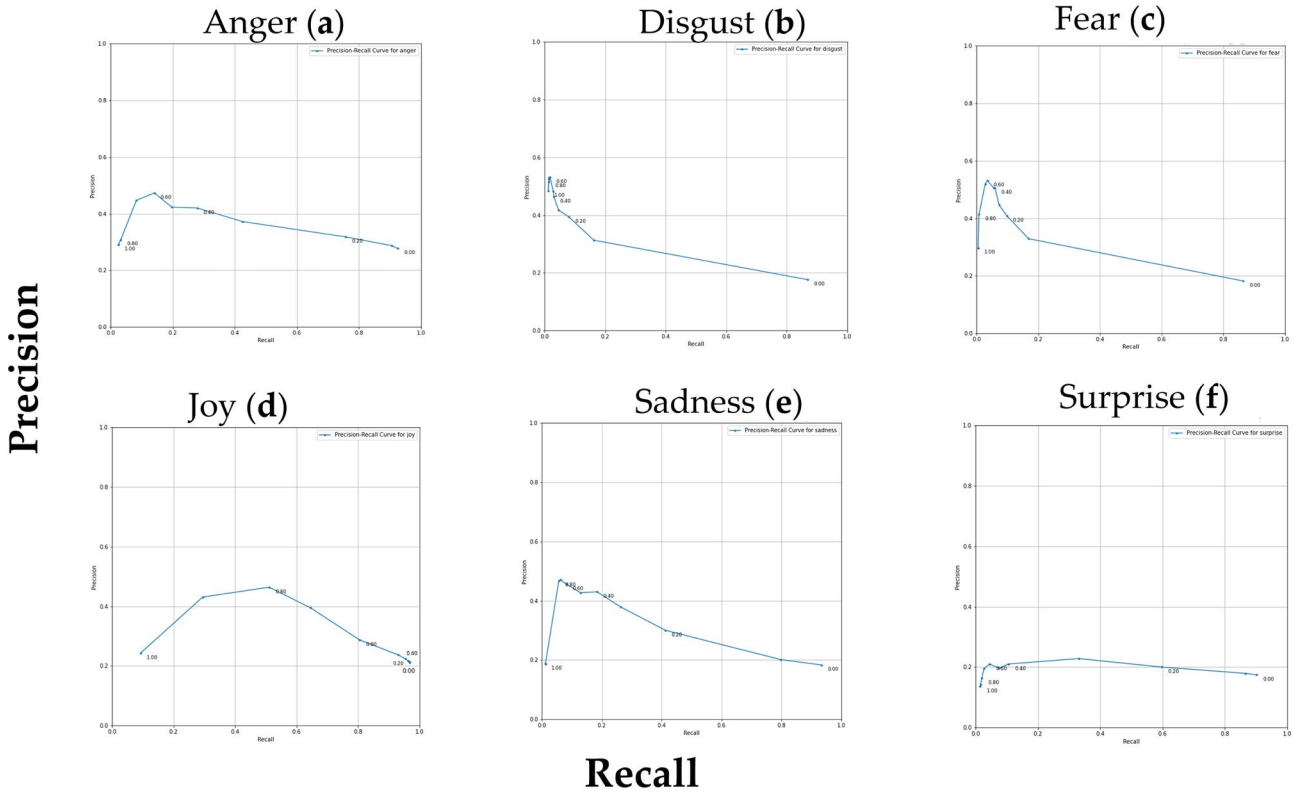


Figure 8. Precision–recall curves for six different emotions analyzed using the GoEmotions distilled lexicon and applied to the XED dataset. Each graphic represents the precision–recall curve for a specific emotion, with points on the curve representing different thresholds. Emotions include (**a**) anger, (**b**) disgust, (**c**) fear, (**d**) joy, (**e**) sadness, and (**f**) surprise.

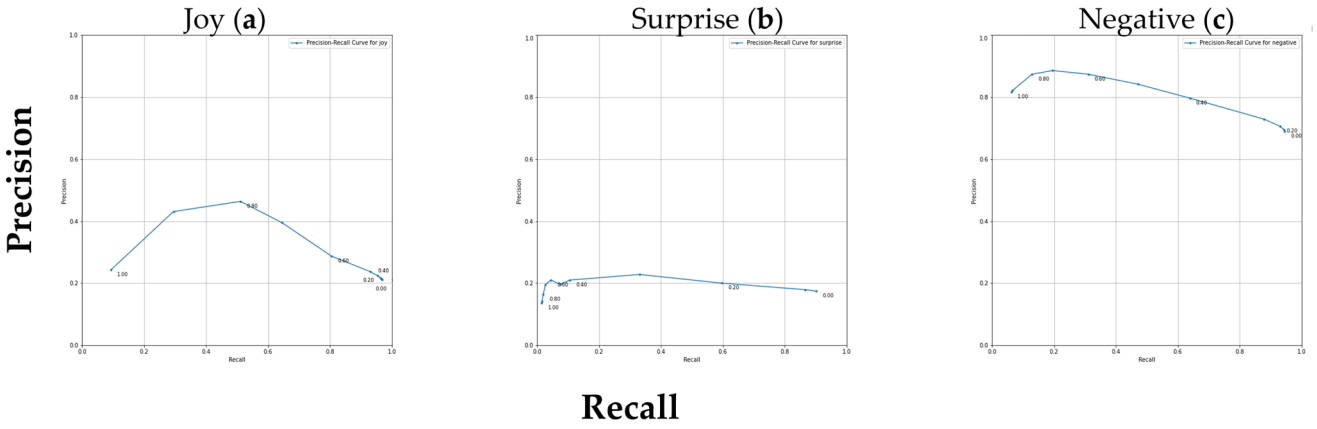


Figure 9. Precision–recall curves for three different emotions analyzed using the GoEmotions Aggregated distilled lexicon and applied to the XED Aggregated dataset. Each graphic represents the precision–recall curve for a specific emotion, with points on the curve representing different thresholds. Emotions include (**a**) joy, (**b**) surprise, and (**c**) negative.

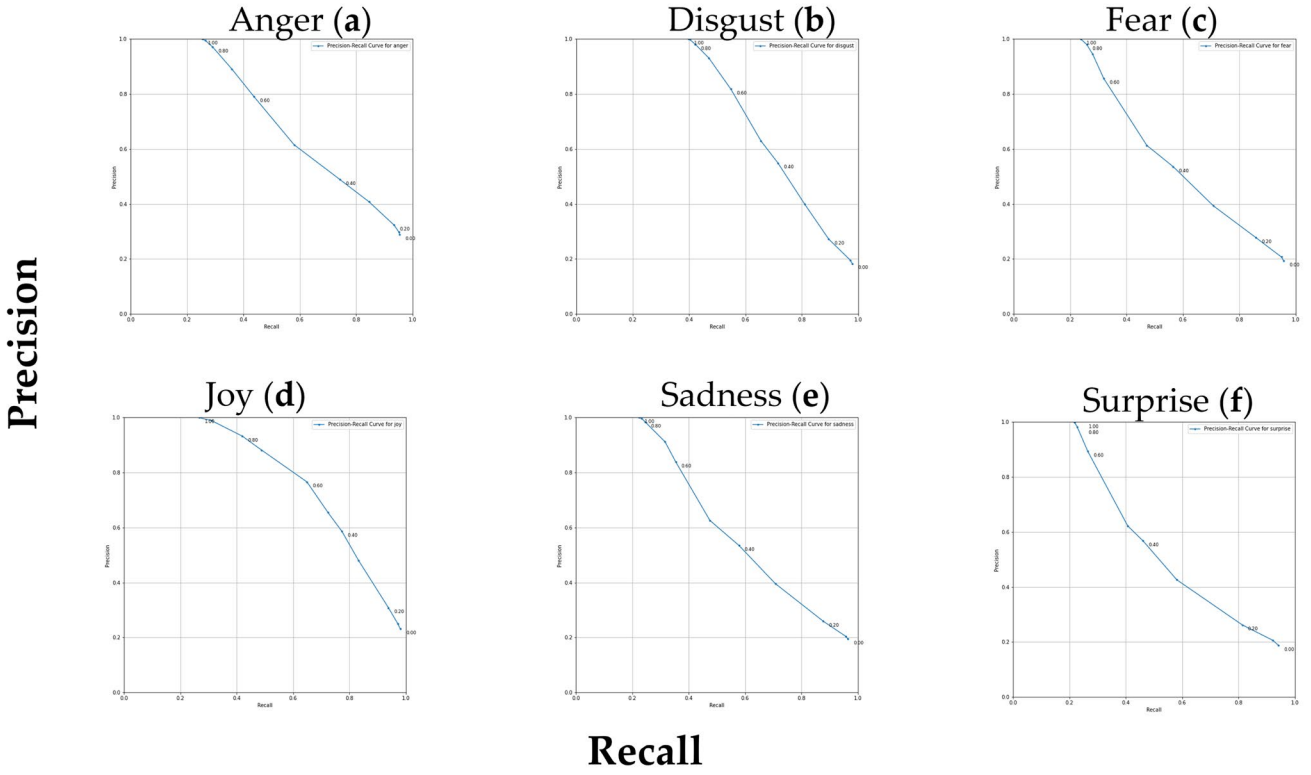


Figure 10. Precision–recall curves for six different emotions analyzed using the XED distilled lexicon and applied to the XED dataset. Each graphic represents the precision–recall curve for a specific emotion, with points on the curve representing different thresholds. Emotions include (**a**) anger, (**b**) disgust, (**c**) fear, (**d**) joy, (**e**) sadness, and (**f**) surprise.

Figure 12 illustrates the precision–recall curves for the XED distilled lexicon applied to the GoEmotions dataset. These curves indicate higher precision and recall for the emotion 'joy' (see Figure 12d), suggesting effective model performance for this specific emotion. Similarly, the performance metrics for aggregated negative emotions shown in Figure 13c also show improvement.

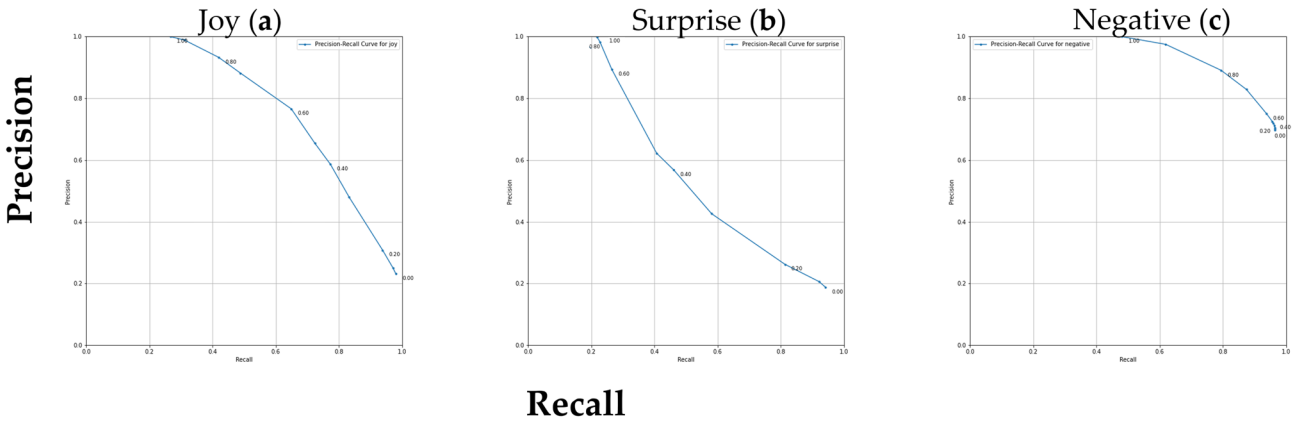


Figure 11. Precision–recall curves for three different emotions analyzed using the XED Aggregated distilled lexicon and applied to the XED Aggregated dataset. Each graphic represents the precision–recall curve for a specific emotion, with points on the curve representing different thresholds. Emotions include (**a**) joy, (**b**) surprise, and (**c**) negative.

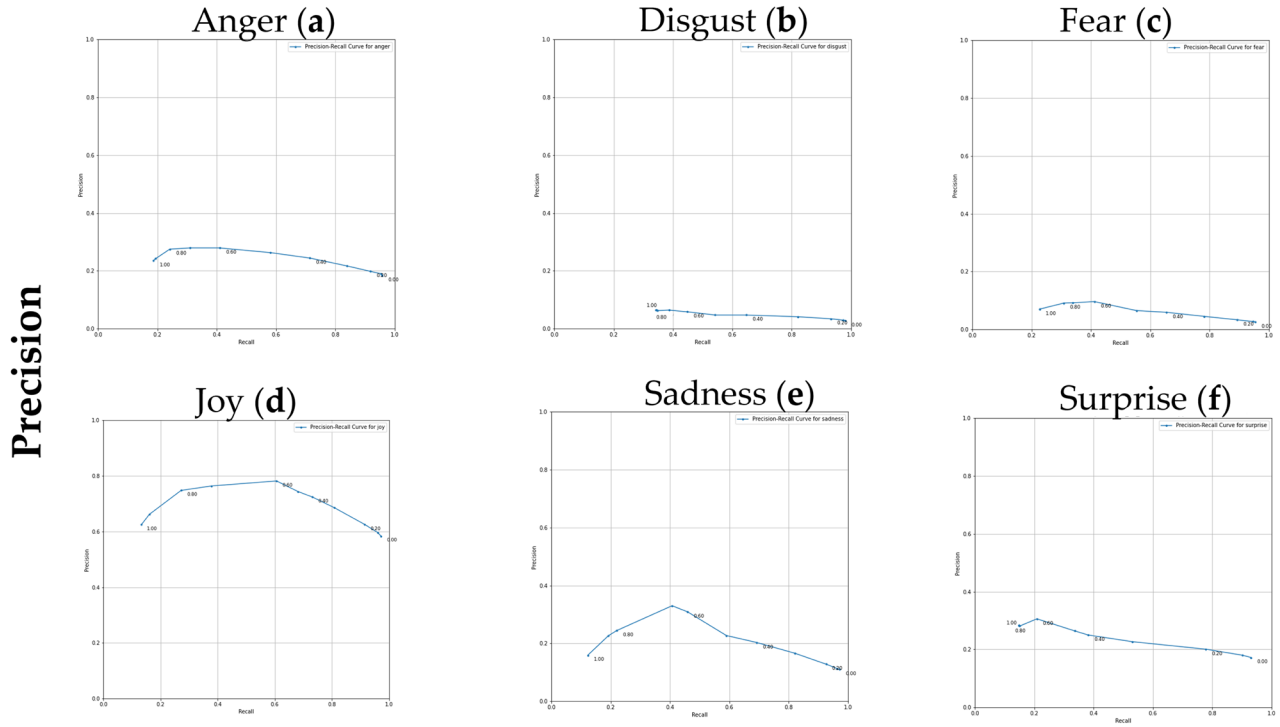


Figure 12. Precision–recall curves for six different emotions analyzed using the XED distilled lexicon and applied to the GoEmotions dataset. Each graphic represents the precision–recall curve for a specific emotion, with points on the curve representing different thresholds. Emotions include (**a**) anger, (**b**) disgust, (**c**) fear, (**d**) joy, (**e**) sadness, and (**f**) surprise.

These precision–recall curves demonstrate that the emotion 'joy' tends to be more accurately identified by the models. A plausible explanation is that 'joy' is the only explicitly positive emotion in the Ekman emotion system, which also categorizes four distinct negative emotions: anger, disgust, fear, and sadness. Additionally, the emotion 'surprise' can be either positive or negative. This distribution could lead to a more straightforward identification process for 'joy', as it does not compete or is related to other positive cate-

gories, unlike the negative emotions, which might share overlapping expressive features, potentially confusing the model.

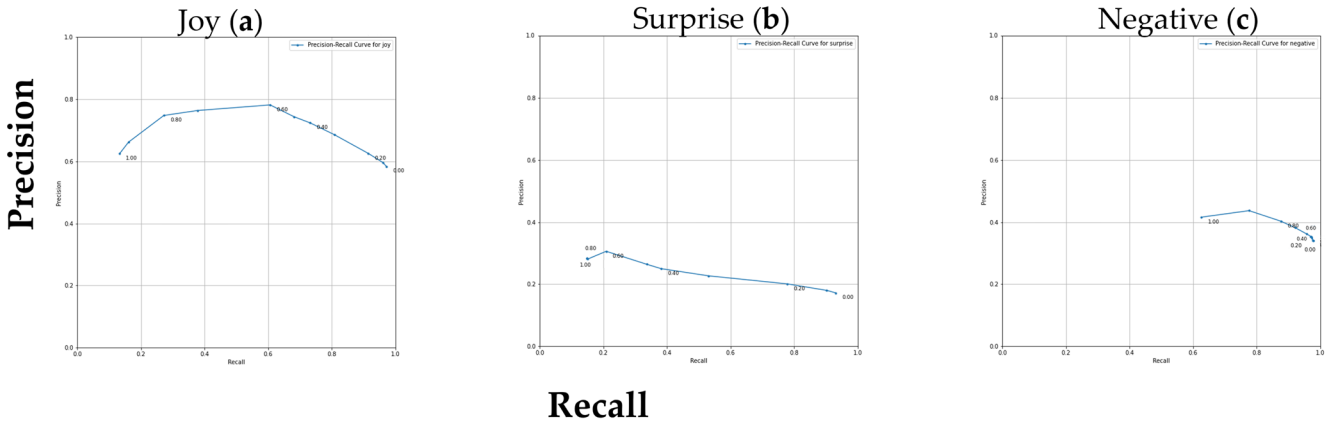


Figure 13. Precision–recall curves for three different emotions analyzed using the XED Aggregated distilled lexicon and applied to the GoEmotions Aggregated dataset. Each graphic represents the precision–recall curve for a specific emotion, with points on the curve representing different thresholds. Emotions include (**a**) joy, (**b**) surprise, and (**c**) negative.

4.3. F1-Score Variants

The following figures comprehensively compare the different F1-score variants—micro, macro, and weighted—used in this study. This analysis aims to illustrate how each metric performs under various conditions and to identify which F1-score variant most effectively captures the nuances of emotion classification across the datasets. The empirical results provide insights into the performance of each metric, highlighting their strengths and limitations in different dataset conditions, principally in terms of class balance and imbalance.

Figure 14a illustrates the overlap between the macro, micro, and weighted F1-scores for the XED lexicon applied to the XED dataset. The maximum values of these metrics are close to each other, indicating a similar level of performance at certain thresholds. However, in the case of the lexicon and dataset with aggregated negative emotions, the macro F1-score consistently shows lower values compared to the micro and weighted F1-scores, which exhibit higher values across the evaluated thresholds (see Figure 14b). Table 4 presents the data from Figure 14 in more detail.

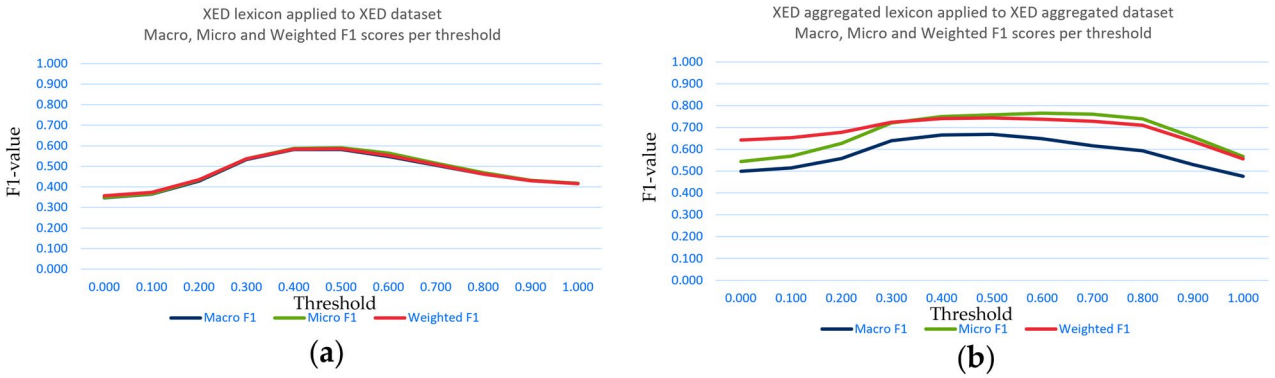


Figure 14. Performance of XED distilled lexicon applied to XED dataset. (**a**) Chart that depicts micro, macro, and weighted F1-scores for the XED distilled lexicon, applied to the XED dataset across various thresholds; (**b**) chart that depicts micro, macro, and weighted f1-scores for the XED Aggregated distilled lexicon, applied to the XED Aggregated dataset across various thresholds.

Table 4. Evaluation metrics for the XED distilled lexicon applied to the XED dataset and for the XED Aggregated distilled lexicon applied to the XED Aggregated dataset.

	XED Distilled Lexicon Applied to XED Dataset			XED Aggregated Distilled Lexicon Applied to XED Aggregated Dataset		
Threshold	**Macro F1**	**Micro F1**	**Weighted F1**	**Macro F1**	**Micro F1**	**Weighted F1**
0.000	0.347	0.349	0.356	0.499	0.545	0.643
0.100	0.365	0.367	0.373	0.515	0.568	0.653
0.200	0.430	0.434	0.435	0.558	0.627	0.677
0.300	0.533	0.537	0.536	0.640	0.720	0.724
0.400	0.582	0.588	0.584	0.666	0.749	0.740
0.500	0.582	0.591	0.585	0.668	0.757	0.743
0.600	0.549	0.565	0.553	0.648	0.765	0.738
0.700	0.506	0.516	0.509	0.616	0.760	0.728
0.800	0.463	0.469	0.464	0.593	0.739	0.710
0.900	0.431	0.433	0.430	0.530	0.656	0.636
1.000	0.417	0.418	0.415	0.476	0.567	0.557

Figure 15a demonstrates variations in performance across different F1-scores when the XED lexicon is applied to the GoEmotions dataset. The macro F1-score is notably lower compared to the micro and weighted F1-scores. Specifically, the macro and the micro F1 do not reach the benchmark value of 0.53, which suggests suboptimal model performance. In contrast, the weighted F1-score indicates an acceptable performance level, at the 0.3 threshold. The XED aggregated lexicon applied to the GoEmotions aggregated dataset has a better performance in the minority class, as the macro F1-score shows higher levels (see Figure 15b). Table 5 present the data shown in Figure 15.

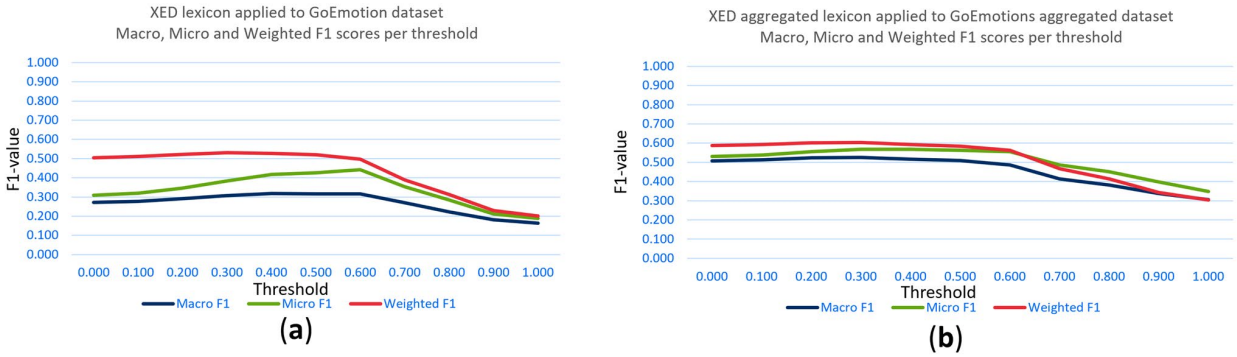


Figure 15. Performance of the XED distilled lexicon applied to the GoEmotions dataset. (**a**) Chart that depicts the micro, macro, and weighted F1-scores for the XED distilled lexicon, applied to the GoEmotions dataset across various thresholds; (**b**) chart that depicts the micro, macro, and weighted F1-scores for the XED Aggregated distilled lexicon, applied to the GoEmotions Aggregated dataset across various thresholds.

Figure 16a for the GoEmotions distilled lexicon applied to the GoEmotions dataset presents a lower macro F1. In the case of the weighted F1, the lower thresholds show a higher value; then, for the higher value thresholds, the micro F1 has higher values (see Table 6). This performance pattern is mirrored in the results from the GoEmotions aggregated lexicon applied to the aggregated GoEmotions dataset, shown in Figure 16b and Table 6. Similar trends across the macro, micro, and weighted F1-scores are observed here. This underscores the impact of lexicon and dataset configuration and how aggregating the negative emotions in one class, as well as having a more balanced dataset, has an impact on the effectiveness of different F1 scoring methods.

Table 5. Evaluation metrics for the XED distilled lexicon applied to the GoEmotions dataset and for the XED Aggregated distilled lexicon applied to the GoEmotions Aggregated dataset.

	XED Distilled Lexicon Applied to the GoEmotions Dataset			XED Aggregated Distilled Lexicon Applied to the GoEmotions Aggregated Dataset		
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
0.600	0.315	0.443	0.497	0.486	0.556	0.563
0.700	0.269	0.353	0.389	0.414	0.487	0.467
0.800	0.221	0.284	0.313	0.382	0.451	0.414
0.900	0.182	0.212	0.229	0.338	0.397	0.342
1.000	0.163	0.188	0.201	0.305	0.348	0.303

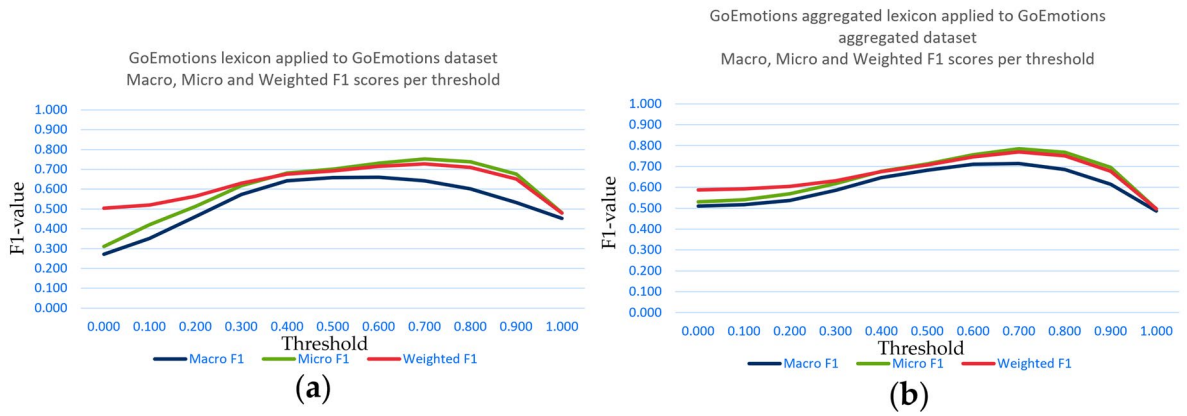


Figure 16. Performance of GoEmotions distilled lexicon applied to GoEmotions dataset. (**a**) Chart that depicts micro, macro, and weighted F1-scores for the GoEmotions distilled lexicon, applied to the GoEmotions dataset across various thresholds; (**b**) chart that depicts micro, macro, and Weighted F1-scores for the GoEmotions Aggregated distilled lexicon, applied to the GoEmotions Aggregated dataset across various thresholds.

Figure 17a evaluates the performance of the GoEmotions lexicon when applied to the XED dataset, revealing that all F1-scores—micro, macro, and weighted—show values below 0.53 (see Table 7). This signifies low model performance. In this case, the micro F1-score outperforms the macro and weighted F1-scores, and none of the three metrics reach the benchmark value of 0.53.

Similarly, when applying the aggregated GoEmotions lexicon to the aggregated XED dataset, the results consistently show a lower macro F1-score that does not reach the threshold of 0.53 (see Figure 17b and Table 7). In this case, the micro and weighted F1-scores are higher and surpass the 0.53 threshold.

The behavior of the micro F1-score, as observed in Figures 14 and 17, reflects its sensitivity to class imbalance. This is noticeable when applying the GoEmotions lexicon to the balanced XED dataset, where the micro F1-score shows higher values compared to other F1 variants. This observation aligns with the characteristic of the micro F1-score of aggregating contributions from all classes, and then reflecting the overall performance more reliably when classes are balanced. The results confirm that the micro F1-score is an appropriate metric for datasets with balanced class distributions.

Table 6. Evaluation metrics for the GoEmotions distilled lexicon applied to the GoEmotions dataset and for the GoEmotions Aggregated distilled lexicon applied to the GoEmotions Aggregated dataset.

	GoEmotions Distilled Lexicon Applied to the GoEmotions Dataset			GoEmotions Aggregated Distilled Lexicon Applied to the GoEmotions Aggregated Dataset		
Threshold	**Macro F1**	**Micro F1**	**Weighted F1**	**Macro F1**	**Micro F1**	**Weighted F1**
0.000	0.272	0.310	0.503	0.510	0.531	0.588
0.100	0.351	0.421	0.520	0.517	0.540	0.593
0.200	0.462	0.514	0.565	0.537	0.569	0.605
0.300	0.573	0.618	0.630	0.586	0.619	0.632
0.400	0.643	0.681	0.677	0.647	0.676	0.676
0.500	0.658	0.700	0.692	0.681	0.713	0.708
0.600	0.660	0.732	0.716	0.710	0.757	0.746
0.700	0.642	0.752	0.728	0.714	0.785	0.770
0.800	0.601	0.738	0.710	0.686	0.768	0.751
0.900	0.532	0.675	0.652	0.614	0.695	0.677
1.000	0.452	0.480	0.480	0.486	0.498	0.497

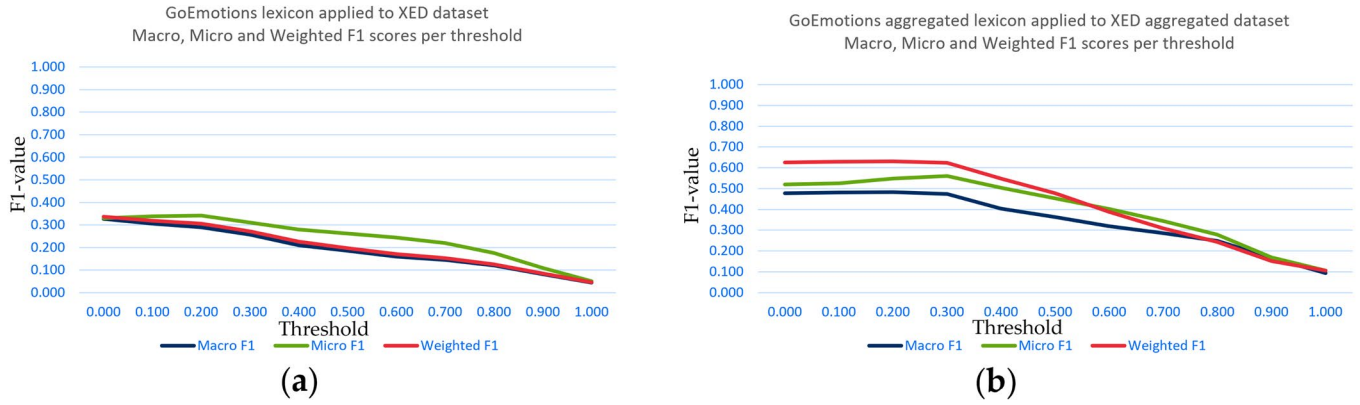


Figure 17. Performance of the GoEmotions distilled lexicon applied to the XED dataset. (**a**) Chart that depicts the micro, macro, and weighted F1-scores for the GoEmotions distilled lexicon, applied to the XED dataset across various thresholds; (**b**) chart that depicts the micro, macro, and Weighted F1-scores for the GoEmotions Aggregated distilled lexicon, applied to the XED Aggregated dataset across various thresholds.

Table 7. Evaluation metrics for the GoEmotions distilled lexicon applied to the XED dataset and for the GoEmotions Aggregated distilled lexicon applied to the XED Aggregated dataset.

	GoEmotions Distilled Lexicon Applied to the XED Dataset			GoEmotions Aggregated Distilled Lexicon Applied to the XED Aggregated Dataset		
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
0.500	0.185	0.262	0.198	0.363	0.453	0.478
0.600	0.159	0.245	0.171	0.319	0.401	0.388
0.700	0.145	0.219	0.154	0.286	0.344	0.308
0.800	0.121	0.177	0.125	0.249	0.278	0.244
0.900	0.081	0.109	0.085	0.167	0.170	0.151
1.000	0.044	0.051	0.046	0.093	0.107	0.106

The weighted F1-score demonstrates that it is sensitive to the presence of more labeled tokens in the lexicons, which can be seen in the lower thresholds, and this might explain the higher values that it shows at lower thresholds. However, when applied to the XED dataset, the weighted F1 performs comparably to the macro F1-score. As can be seen in Table 2, the XED dataset is balanced. This trend in the weighted F1-score changes with the aggregated XED dataset, where negative emotions are combined into a single class, thus skewing the class balance. Under these conditions, at lower thresholds, the weighted F1 shows better performance. It is also a metric that is responsive to changes in class balance and, in this case, word availability.

5. Conclusions

This study confirms findings from the existing literature on the importance of selecting appropriate evaluation metrics, particularly different types of F1-scores, in multilabel emotion classification tasks. Specifically, it highlights that relying on a single metric does not provide a comprehensive picture of classifier performance.

In this paper, the XED distilled lexicon applied to the GoEmotions dataset reached the benchmark value of 0.53 at a threshold of 0.3. This was the only experiment among the non-aggregated datasets to achieve this score while applying the lexicon to a different dataset than the one from it was originally derived.

It was also found that the configuration of the emotion system used in the annotated dataset matters. This was demonstrated by aggregating emotions into fewer categories, which improved model performance metrics, especially for macro F1-scores. In this case, the XED Aggregated distilled lexicon applied to the GoEmotions Aggregated dataset and the GoEmotions Aggregated distilled lexicon applied to the XED Aggregated dataset both achieved results above 0.53 in the micro and weighted F1 for certain thresholds.

The main limitation of this study is that only the Ekman emotion classification system was used. Furthermore, only unigrams were considered when using the distilled lexicons, which means that the context of the messages was not taken into consideration, impacting the classifier's performance.

For future research, we suggest further exploring the influence of class imbalance and category aggregation on emotion classification models. Specifically, different emotion systems, such as Plutchik's, dimensional, or componential models, should be used and compared with the results obtained from using Ekman's system.

In conclusion, we consider that presenting the three F1-score variants—weighted, macro, and micro F1-scores—when evaluating the performance of a classifier offers a more comprehensive picture of the model in imbalanced scenarios than when relying on a single metric. We recommend these metrics to be used together with the F1-scores per class for multilabel classification tasks, as this could avoid the issues caused by averaging them and assigning weights.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app14219863/s1.

Author Contributions: Conceptualization, M.C.H.L., J.B. and J.S.; data curation, M.C.H.L.; formal analysis, M.C.H.L.; funding acquisition, J.B. and J.S.; investigation, M.C.H.L.; methodology, M.C.H.L., J.B. and J.S.; project administration, M.C.H.L.; resources, M.C.H.L., J.B. and J.S.; software, M.C.H.L.; supervision, J.B. and J.S.; validation, M.C.H.L., J.B. and J.S.; visualization, M.C.H.L.; writing—original draft, M.C.H.L.; writing—review and editing, M.C.H.L., J.B. and J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on Hugging Face. The GoEmotions dataset was retrieved from https://huggingface.co/datasets/google-research-datasets/go_emotions, and it is described in Demszky et al. (2020) [37], accessed on 24 October 2023. The XED dataset was obtained from https://huggingface.co/datasets/Helsinki-NLP/xed_en_fi, and it is detailed in Öhman et al. (2020) [36], (accessed on 24 October 2023). These datasets are publicly available and have been used in previous research for sentiment analysis and emotion detection.

Acknowledgments: For this paper, Grammarly was used for grammar correction, and ChatGPT was employed to improve the clarity and flow of the text, for the basic debugging of code, and for brainstorming. However, no generative AI was involved in producing original research content or findings. We would also like to thank Tyché Perkisas for his invaluable help and support throughout this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Alpaydin, E. *Machine Learning: The New AI*; MIT Press: Cambridge, MA, USA, 2016.
2. Abbasi, B.; Goldenholz, D.M. Machine learning applications in epilepsy. *Epilepsia* **2019**, *60*, 2037–2047. [CrossRef] [PubMed]
3. Ford, V.; Siraj, A. Applications of Machine Learning in Cyber Security. In Proceedings of the 27th International Conference on Computer Applications in Industry and Engineering, CAINE 2014, New Orleans, LA, USA, 13–15 October 2014.
4. Rajkomar, A.; Dean, J.; Kohane, I. Machine Learning in Medicine. *N. Engl. J. Med.* **2019**, *380*, 1347–1358. [CrossRef] [PubMed]
5. Kelleher, J.; Mac Namee, B.; D'Arcy, A. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*; MIT Press: Cambridge, MA, USA, 2015.
6. Rainio, O.; Teuho, J.; Klen, R. Evaluation metrics and statistical tests for machine learning. *Sci. Rep.* **2024**, *14*, 6086. [CrossRef]
7. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
8. Takahashi, K.; Yamamoto, K.; Kuchiba, A.; Koyama, T. Confidence interval for micro-averaged F_1 and macro-averaged F_1 scores. *Appl. Intell.* **2022**, *52*, 4961–4972. [CrossRef]
9. Ting, K.M. Precision and Recall. In *Encyclopedia of Machine Learning*; Sammut, C., Webb, G.I., Eds.; Springer: Boston, MA, USA, 2010; p. 781.
10. Rijsbergen, C.J.V. *Information Retrieval*; Butterworths: Newton, MA, USA, 1979; p. 208.
11. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1-score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef] [PubMed]
12. Miao, J.; Zhu, W. Precision–recall curve (PRC) classification trees. *Evol. Intell.* **2021**, *15*, 1545–1569. [CrossRef]
13. Poudel, S.; Bikdash, M. Optimal Dependence of Performance and Efficiency of Collaborative Filtering on Random Stratified Subsampling. *Big Data Min. Anal.* **2022**, *5*, 192–205. [CrossRef]
14. Poudel, S. A study of disease diagnosis using machine learning. *Med. Sci. Forum* **2022**, *10*, 8. [CrossRef]
15. Ramezani, A.; El-henawy, I.; Zidan, M.; Othman, M. BERT-CNN: A Deep Learning Model for Detecting Emotions from Text. *Comput. Mater. Cont.* **2021**, *71*, 2943–2961. [CrossRef]
16. Shelke, N.; Chaudhury, S.; Chakrabarti, S.; Bangare, S.L.; Yogapriya, G.; Pandey, P. An efficient way of text-based emotion analysis from social media using LRA-DNN. *Neurosci. Inf.* **2022**, *2*, 100048. [CrossRef]
17. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 27. [CrossRef]
18. Bekkar, M.; Djemaa, H.K.; Alitouche, T.A. Evaluation Measures for Models Assessment over Imbalanced Data Sets. *J. Inf. Eng. Appl.* **2013**, *3*, 27–38.
19. Powers, D. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *J. Mach. Learn. Technol.* **2011**, *2*, 2229–3981. [CrossRef]
20. Gu, Q.; Zhu, L.; Cai, Z. Evaluation Measures of the Classification Performance of Imbalanced Data Sets. In Proceedings of the Computational Intelligence and Intelligent Systems: 4th International Symposium, ISICA 2009, Huangshi, China, 23–25 October 2009; Springer: Berlin/Heidelberg, Germany, 2009; pp. 461–471.
21. Lango, M. Tackling the Problem of Class Imbalance in Multi-class Sentiment Classification: An Experimental Study. *Found. Comput. Decis. Sci.* **2019**, *44*, 151–178. [CrossRef]
22. Xu, R.F.; Chen, T.; Xia, Y.Q.; Lu, Q.; Liu, B.; Wang, X. Word Embedding Composition for Data Imbalances in Sentiment and Emotion Classification. *Cogn. Comput.* **2015**, *7*, 226–240. [CrossRef]
23. Akosa, J.S. Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data. In Proceedings of the SAS Global Forum 2017 Conference, Orlando, FL, USA, 2–5 April 2017.
24. Flach, P.A.; Kull, M. Precision-Recall-Gain Curves: PR Analysis Done Right. In Proceedings of the Advances in Neural Information Processing Systems 28 (Nips 2015), Montreal, QC, Canada, 7–12 December 2015; Volume 28.
25. Heydarian, M.; Doyle, T.E.; Samavi, R. MLCM: Multi-Label Confusion Matrix. *IEEE Access* **2022**, *10*, 19083–19095. [CrossRef]
26. Chicco, D.; Jurman, G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Min.* **2023**, *16*, 4. [CrossRef]

27. Singh, S.; Khim, J. Optimal binary classification beyond accuracy. In Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; Article 1325. pp. 18226–18240.
28. Perkisas, T. On Dedicated Anti-Cybercrime Mailboxes Within Financial Institutions: Characterization & Evolution. Ph.D. Thesis, University of Antwerp, Antwerp, Belgium, 2024. [CrossRef]
29. Nandwani, P.; Verma, R. A review on sentiment analysis and emotion detection from text. *Soc. Netw. Anal. Min.* **2021**, *11*, 81. [CrossRef]
30. Munezero, M.; Suero Montero, C.; Sutinen, E.; Pajunen, J. Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text. *IEEE Trans. Affect. Comput.* **2014**, *5*, 101–111. [CrossRef]
31. Seyeditabari, A.; Tabari, N.; Zadrozny, W. Emotion Detection in Text: A Review. *arXiv* **2018**, arXiv:1806.00674.
32. Wortman, B.; Wang, J.Z. HICEM: A High-Coverage Emotion Model for Artificial Emotional Intelligence. *IEEE Trans. Affect. Comput.* **2024**, *15*, 1136–1152. [CrossRef]
33. Tracy, J.; Randles, D. Four Models of Basic Emotions: A Review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. *Emot. Rev.* **2011**, *3*, 397–405. [CrossRef]
34. Zhang, F.; Chen, J.; Tang, Q.; Tian, Y. Evaluation of emotion classification schemes in social media text: An annotation-based approach. *BMC Psychol.* **2024**, *12*, 503. [CrossRef] [PubMed]
35. Hugging Face. The AI Community Building the Future. Datasets. Available online: https://huggingface.co/datasets (accessed on 24 October 2023).
36. Öhman, E.; Pàmies, M.; Kajava, K.; Tiedemann, J. XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 6542–6552.
37. 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), Online, 5–10 July 2020; pp. 4040–4054. [CrossRef]
38. Erickson, B.J.; Kitamura, F. Magician's Corner: 9. Performance Metrics for Machine Learning Models. *Radiol. Artif. Intell.* **2021**, *3*, e200126. [CrossRef]
39. Gupta, A.; Anand, A.; Hasija, Y. Recall-based Machine Learning approach for early detection of Cervical Cancer. In Proceedings of the 2021 6th International Conference for Convergence in Technology (I2CT), Maharashtra, India, 2–4 April 2021; pp. 1–5.
40. Olek, M. About Evaluation of F1-score for RECENT Relation Extraction System. *arXiv* **2023**, arXiv:2305.09410.
41. Strapparava, C.; Mihalcea, R. SemEval-2007 Task 14: Affective Text. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, 23–24 June 2007; pp. 70–74.
42. Huang, C.; Trabelsi, A.; Zaïane, O. Seq2Emo for Multi-label Emotion Classification Based on Latent Variable Chains Transformation. *arXiv* **2019**, arXiv:1911.02147.
43. Abdul-Mageed, M.; Ungar, L. EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 718–728.
44. Liu, C.; Osama, M.; De Andrade, A. DENS: A Dataset for Multi-class Emotion Analysis. *arXiv* **2019**, arXiv:1910.11769; pp. 6293–6298, 6293–6298.
45. Davis, J.; Goadrich, M. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning—ICML'06, Pittsburgh, PA, USA, 25–29 June 2006; pp. 233–240.
46. Andrikakis, E.; Perikos, I.; Paraskevas, M.; Hatzilygeroudis, I. Text Analysis and Recognition of Emotional Content Using Deep Learning Methods and BERT. In Proceedings of the 2023 IEEE/ACIS 23rd International Conference on Computer and Information Science (ICIS), Wuxi, China, 23–25 June 2023; pp. 37–42.
47. Ekman, P. An Argument for Basic Emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [CrossRef]
48. Wang, H.; Xu, Q.; Zhou, L. Large unbalanced credit scoring using Lasso-logistic regression ensemble. *PLoS ONE* **2015**, *10*, e0117844. [CrossRef]
49. Fang, J. The role of data imbalance bias in the prediction of protein stability change upon mutation. *PLoS ONE* **2023**, *18*, e0283727. [CrossRef] [PubMed]
50. SPACY Industrial. Strength Natural Language Processing in Python. Available online: https://spacy.io/ (accessed on 20 April 2024).
51. Kaur, J.; Buttar, P. A Systematic Review on Stopword Removal Algorithms. *Inf. Process. Manag.* **2018**, *4*, 207–210.
52. Mandl, T.; Modha, S.; Majumder, P.; Patel, D.; Dave, M.; Mandlia, C.; Patel, A. Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation (Fire 2019), Kolkata, India, 12–15 December 2019; pp. 14–17. [CrossRef]
53. Opitz, J.; Burst, S. Macro F1 and Macro F1. *arXiv* **2019**, arXiv:1911.03347.
54. Harbecke, D.; Hennig, L.; Alt, C. Why only Micro-F1? Class Weighting of Measures for Relation Classification. *arXiv* **2022**, arXiv:2205.09460.
55. Sun, Y.M.; Wong, A.K.C.; Kamel, M.S. Classification of Imbalanced Data: A Review. *Int. J. Pattern Recognit. Artif. Intell.* **2009**, *23*, 687–719. [CrossRef]

56. Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef]
57. Joshi, K.; Kumar, S.; Rawat, J.; Kumari, A.; Gupta, A.; Sharma, N. Fraud App Detection of Google Play Store Apps Using Decision Tree. In Proceedings of the 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), Pradesh, India, 23–25 February 2022; pp. 243–246.
58. Coden, A.; Savova, G.; Sominsky, I.; Tanenblatt, M.; Masanz, J.; Schuler, K.; Cooper, J.; Guan, W.; de Groen, P.C. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. *J. Biomed. Inform.* **2009**, *42*, 937–949. [CrossRef]
59. Kane, A.; Patankar, S.; Khose, S.; Kirtane, N. Transformer based ensemble for emotion detection. In Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, Dublin, Ireland, 26 May 2022; pp. 250–254. [CrossRef]
60. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
61. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv* **2016**, arXiv:1606.05250.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.