University of Antwerp
Faculty of Business and Economics

DEPARTMENT OF ENGINEERING MANAGEMENT

Evaluating Performance Metrics in Emotion Lexicon Distillation: A Focus on F1 Scores

Maria Cristina Hinojosa-Lee, Johan Braet & Johan Springael

UNIVERSITY OF ANTWERP
Faculty of Business and Economics
City Campus
Prinsstraat 13
B-2000 Antwerp
www.uantwerpen.be

FACULTY OF BUSINESS AND ECONOMICS

DEPARTMENT OF ENGINEERING MANAGEMENT

Evaluating Performance Metrics in Emotion Lexicon Distillation: A Focus on F1 Scores

Maria Cristina Hinojosa-Lee, Johan Braet & Johan Springael

RESEARCH PAPER 2024-002
JULY 2024

University of Antwerp, City Campus, Prinsstraat 13, B-2000 Antwerp, Belgium
Research Administration
e-mail: joeri.nys@uantwerpen.be

The research papers from the Faculty of Business and Economics are also available at www.repec.org (Research Papers in Economics - RePEc)

D/2024/1169/002

Evaluating Performance Metrics in Emotion Lexicon Distillation: A Focus on F1 Scores

M.C. Hinojosa-Lee, J. Braet, J. Springael

Department of Engineering Management, University of Antwerp

Abstract

This study compares various F1 score variants—micro, macro, and weighted—to evaluate the performance of lexicon distillation. The datasets used for this are the multilabel emotion annotated datasets XED and GoEmotions. The aim of this paper is to understand the effects of class imbalance on the F1 score. 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 to the ground truth. Then, the F1 score variants—micro, macro, and weighted—were calculated and 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. This contribution aims to guide researchers in selecting suitable metrics for emotion analysis classification tasks.

Keywords: emotion analysis, multi label dataset, lexicon, annotated datasets, F1-score, performance metrics

1. Introduction

Machine learning has become an important technology used in diverse sectors, transforming how data is analyzed and utilized (Alpaydin, 2016; Abbasi & Goldenholz, 2019). 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 (Rajkomar et al., 2019; Ford & Siraj, 2014). The widespread application of machine learning techniques increases the need for new researchers to understand basic statistical principles (Rainio et al., 2024). This necessity underscores the importance of correctly applying statistical methods and selecting appropriate metrics to evaluate the performance of models and algorithms (Rainio et al., 2024).

Metrics like precision, recall, and the F1 score, have their origin in disciplines reliant on empirical evidence, including clinical trials, information retrieval, and behavioral research (Sokolova & Lapalme, 2009; Takahashi et al., 2021). These metrics originated in Information Extraction and Information Retrieval and have been adapted to evaluate machine learning algorithms (Sokolova & Lapalme, 2009; Takahashi et al., 2021). 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 (Ting, 2011).

The consequences of mistakes at evaluating classifiers can be particularly severe in fields such as healthcare, where the costs associated with false negatives (FN) and false positives (FP) are significant. For instance, a false negative in a disease diagnosis could result in delaying treatment, while a false positive could lead to unnecessary treatments. This highlights the importance of verifying the reliability of machine learning predictions through appropriate metrics (Alpaydin, 2016). Poudel (2022) explores this by affirming that diagnosing a disease is a classification task evaluated using performance metrics such as accuracy, precision, recall, and F1 score (Poudel & Bikdash, 2022; Poudel, 2022). While these four evaluation metrics are commonly used to assess classifiers, researchers often select only one metric, with accuracy being the preferred choice (Poudel, 2020). However, given the cost sensitivity of FN and FP, it is crucial 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 (Johnson & Khoshgoftaar, 2019). When classes are imbalanced, accuracy may be misleading, presenting high scores that do not represent good performance (Johnson & Khoshgoftaar, 2019; Poudel, 2020). Therefore, when working with an imbalanced dataset, it is recommended to use multiple performance metrics (Poudel, 2020). 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 (Powers, 2011; Sokolova & Lapalme, 2009).

Sentiment and Emotion Analysis frequently encounters the challenge of class imbalance, representing a key obstacle to improving sentiment and emotion classifiers (Xu et al., 2015; Lango, 2019). 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 (Akosa, 2017; Lango, 2019). This issue is evident with the use of accuracy in the presence of 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 (Akosa, 2017).

Despite the documented challenges posed by class imbalance and the common use of precision, recall, and the F1 score, there is a lack of research comparing different variants of the F1 score (micro, macro, and weighted) in the context of lexicon-based emotion analysis. This gap is critical because emotion analysis often involves imbalanced datasets, and addressing it can help in selecting the right evaluation metric to assess a classification model's performance.

Considering this, the goal of this study is to compare the effectiveness of micro, macro, and weighted F1 metrics in reflecting the performance of emotion classification models. The Methodology section of this paper includes the experiment setup, the datasets used, and detailed explanation over the performance metrics to compare. The Results section consists of a comparison between the different types of F1 scores.

2. Methodology

This study used the GoEmotions and XED datasets from the Hugging Face repository (Hugging Face, 2023). The selection of these datasets was motivated by their broad spectrum of vocabulary and topics derived from nonsocial media sources. This choice helps to avoid the potential bias and limitations associated with Twitter based datasets, which are too specific and may not generalize well across different contexts that are less specific such as emails (Öhman et al., 2020). The following subsections detail the methodology employed.

2.1 Datasets

The GoEmotions multilabel dataset, published in 2020, consists of 58000 manually emotion annotated English comments from Reddit (Demszky et al., 2020). It is a multilabel dataset that initially included 27 emotional labels plus a neutral category. The authors curated the messages to reduce profanity and harmful content. 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. According to Demszky et al. (2020),

the 27 emotions are hierarchically grouped into Ekman's six basic emotions: anger, disgust, fear, joy, sadness, and surprise. For this study, we used the Ekman level of emotion classification, excluding the neutral category, resulting in a dataset of 38237 messages (Demszky et al., 2020). The neutral class was not considered for this research. This is because as noted in the source paper, it did not include messages with no emotion or with all emotions present, but messages that were difficult to classify as containing a specific emotion. Figure 1 shows how the 27 emotions were clustered into the six basic emotions system proposed by Ekman.

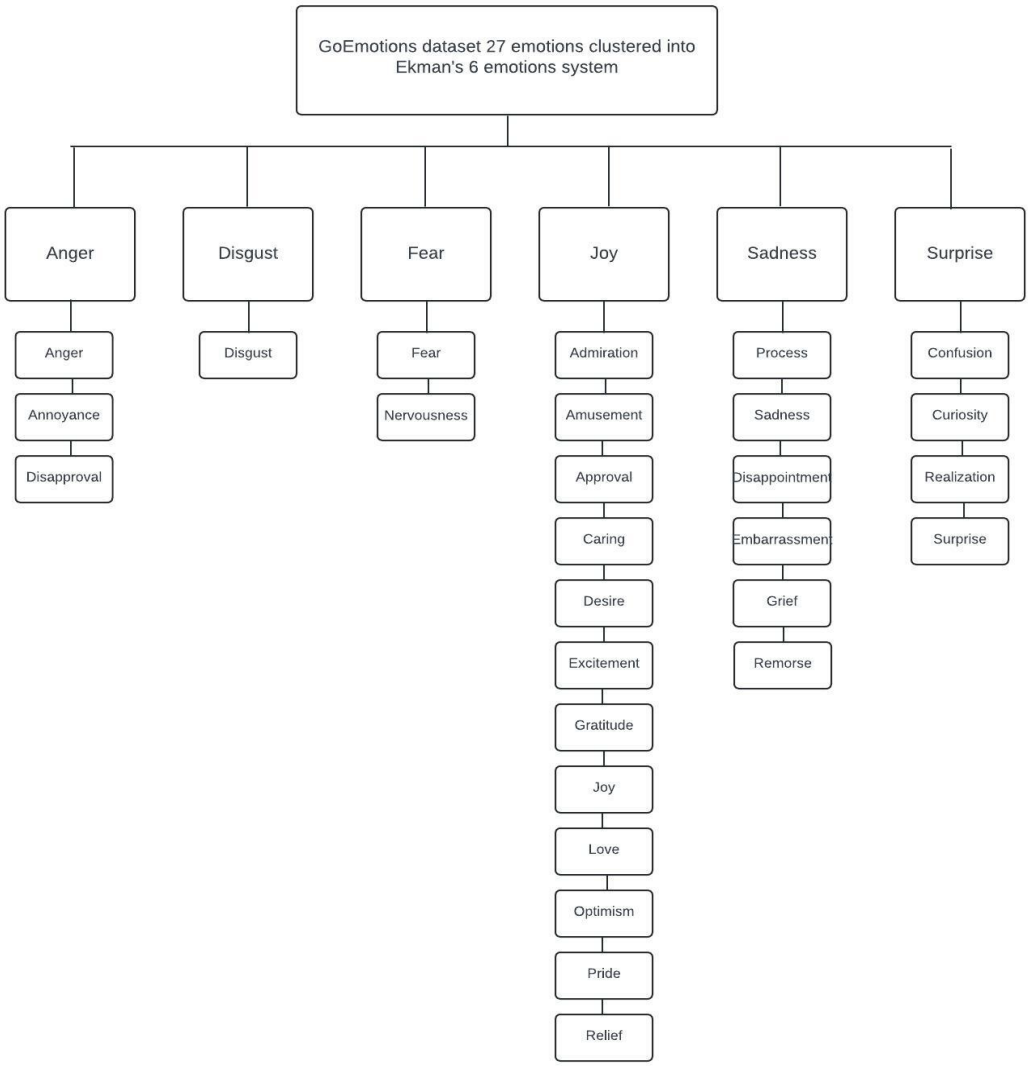


FIGURE 1: DIAGRAM SHOWING THE DIFFERENT 27 EMOTIONS USED IN THE GOEMOTIONS DATASET AND HOW THEY GET CLUSTERED INTO EKMAN'S 6 EMOTIONS (DEMSZKY ET AL., 2020).

The XED dataset includes 25000 Finnish and 30000 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 carried out the annotations, requiring a consensus from at least

two of three annotators for each labeled movie subtitle (Öhman et al., 2020). For this experiment, the emotion labels 'anticipation,' 'trust,' and 'neutral' were excluded. The emotions considered were those that are part of the Ekman emotion theory, a classification also used in the GoEmotions dataset: anger, disgust, fear, joy, sadness, and surprise (Ekman, 1992; Öhman et al., 2020). Öhman et al. (2020) affirm that the granularity of their annotations is at the sentence level; however, they also mention that the shortest movie subtitle they have is "!" and the longest subtitle includes three sentences. 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. This considerations led to the use of 13682 English messages (see Table 1).

TABLE 1: 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	Origin	Number of messages	Emotions					
			Anger	Disgust	Fear	Joy	Sadness	Surprise
GoEmotions	Reddit forum	38237	7021	1012	929	21730	4030	6668
XED	Subtitles from movies	13682	3828	2317	2439	2832	2464	2442

A common problem with data is that it can be skewed, with the majority of examples pertaining to one class (Wang et al., 2015; Fang, 2023; Akosa, 2017). When this occurs, the dataset is said to be imbalanced (Wang et al., 2015; Fang, 2023). A balanced dataset has a similar number of cases in each class (Fang, 2023). 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 (Fang, 2023).

Table 1 shows that, with the exception of 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 those labeled as 'fear.' This imbalance may stem from using Ekman's emotion classification, where 'joy' is the only positive emotion included (Demszky et al., 2020). 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 2 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 2: 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	38237	12942	21730	6668
XED Aggregated	Subtitles from movies	13682	9425	2832	2442

2.2 Data Preprocessing

Both datasets underwent identical preprocessing to standardize the text data for consistent analysis. Using the SpaCy library (Honnibal & Montani, 2023), stopwords and non-alphanumeric characters were removed from the messages, and the text was lowercased. The idea behind removing stopwords is that they are common words that might not convey information and may reduce accuracy and performance if they are not eliminated (Buttar, 2018).

2.3 Lexicon Distillation Process

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

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. 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 that emotion.
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. 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 one 1: meaning association.

Figure 2 shows a flow chart that describes the process of distilling a lexicon from an annotated dataset.

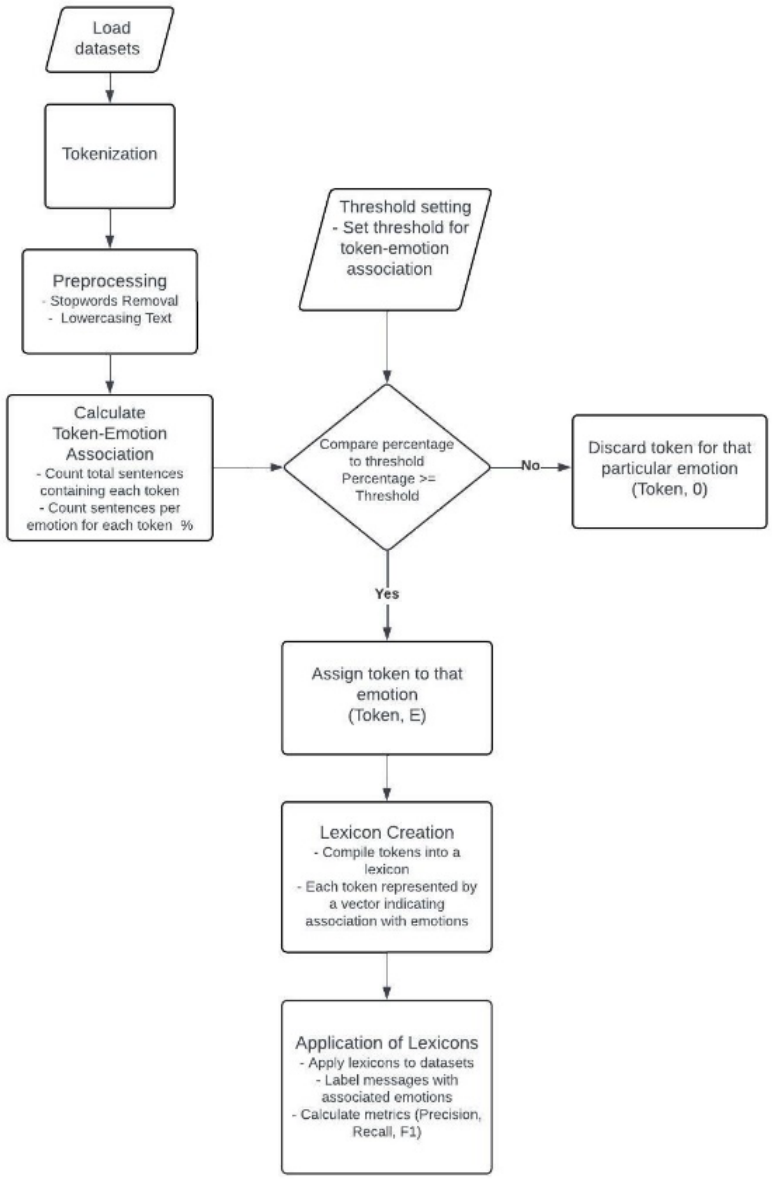


FIGURE 2: FLOW CHART DESCRIBING THE PROCESS OF DISTILLING A LEXICON FROM AN ANNOTATED DATASET

2.4 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 positives (TP) for each class, and then averaging the outcomes (Mandl et al., 2019). The weighted F1 score is used when there are imbalanced classes and shows a bias towards the majority class (Mandl et al., 2019; Sokolova & Lapalme, 2009). 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 (Mandl et al., 2019). Consequently, the model is penalized if it performs poorly on the minority class (Opitz & Burst, 2019; Mandl et al., 2019). The micro F1 score is calculated over the entire dataset, this means that it disregards class membership (Harbecke et al., 2022).

The process of applying the lexicon involves scanning each message in the datasets to find words that are present in the lexicon (see Figure 3). Each message is labeled with the emotions associated with the tokens in both the text and the lexicon. If the tokens in a message are associated with different emotions, or if a single token is associated with multiple emotions, the message is labeled to reflect all detected emotions. The labeling is done using a vector for each message, where each component corresponds to one of the emotions (six for the whole model or three for the aggregated model). The vectors are filled with binary values (0 or 1), indicating the absence or presence of each emotion (Figure 3).

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 (Sokolova & Lapalme, 2009; Sun et al., 2009). It is divided into four segments:

1. Number or percentage of examples that are correctly classified as part of a class (true positives or TP).
2. Number or percentage of examples that are correctly classified as not being part of a class (true negatives or TN).
3. Number or percentage of examples that are incorrectly classified as being part of a class (false positives or FP).
4. Number or percentage of examples that are incorrectly classified as not being part of a class (False negatives or FN).

The goal of calculating the confusion matrixes is to assess how well the model could classify the messages. At the same time, the values that form these matrices are the basis for calculating the F1 measures under varying conditions (Erickson & Kitamura, 2021). The structure of a confusion matrix is shown in Table 3.

TABLE 3: CONFUSION MATRIX STRUCTURE (SUN ET AL., 2009)

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		

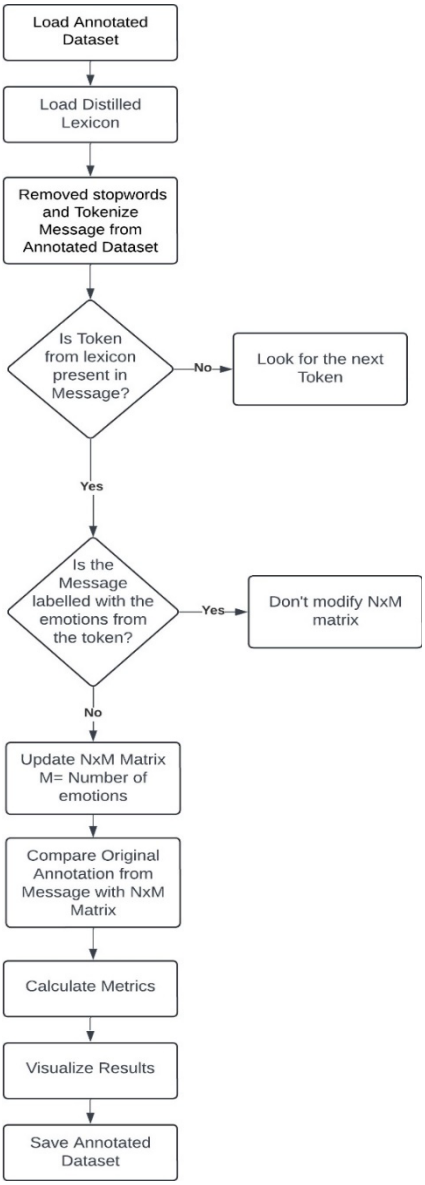


FIGURE 3: FLOW CHART DESCRIBING THE PROCESS OF APPLYING A DISTILLED LEXICON TO AN ANNOTATED DATASET

2.5 Evaluation metrics

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

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 (Erickson & Kitamura, 2021; Gupta et al., 2021). Precision (Equation 2) is the fraction of correctly predicted positive cases from all cases predicted as positive (Erickson & Kitamura, 2021; Gupta et al., 2021). The F1 score, also known as the Dice similarity coefficient, is the harmonic mean of precision and recall, providing a balance between the two (Erickson & Kitamura, 2021; Gupta et al., 2021; Akosa, 2017). 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.

EQUATION 1

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

EQUATION 2

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

Where TP are the true positives, FP are the false positives, and FN represents the false negatives.

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

EQUATION 3

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

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

EQUATION 4

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

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:

EQUATION 5

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

Where w_i is the proportion of true instances for label i. The weighted F1 score takes into account the support or weight of each class, w_i (Equation 6)

EQUATION 6

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

Where TP_i and FN_i are the true positives and the false negatives for class i, 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 evaluating the performance of emotion analysis models (see Table 4), utilizing emotion-annotated datasets ranging from news headlines to book passages, movie subtitles, and messages from Twitter and Reddit (Strapparava & Mihalcea, 2007; Abdul-Mageed & Ungar, 2017; Huang et al., 2019; Liu et al., 2019; Öhman et al., 2020). It is important to note that some researchers still rely solely on accuracy (Strapparava & Mihalcea, 2007; Abdul-Mageed & Ungar, 2017) or the micro F1 score (Huang et al., 2019). Using these measures to evaluate performance can result in overly optimistic outcomes or complicate comparisons across different models and datasets (Davis & Goadrich, 2006; Öhman et al., 2020; Andrikakis et al., 2023).

TABLE 4: EVALUATION METRICS USED IN EMOTION-ANNOTATED DATASETS. THE SYMBOL μ INDICATES THAT THE MICRO F1 SCORE WAS USED, INSTEAD OF THE MACRO F1 SCORE (ÖHMAN ET AL., 2020).

Study	Source	Cat	Multi	Macro F1	Accuracy	Balanced?
Abdul-Mageed and Ungar (2017)	Twitter	8	No	N/A	95.68%	N
Abdul-Mageed and Ungar (2017)	Twitter	24	No	87.47%	N/A	N
Samy et al. (2018)	Twitter	11+neu	Yes	64.8%	53.2%	N
Liu et al. (2019)	Books etc.	8+neu	No	60.4%μ	N/A	Y Only after dropping disgust and surprise.
Demszky et al. (2020)	Reddit	27	Yes	46%	N/A	N
Demszky et al. (2020)	Reddit	6	No	64%	N/A	N
XED (English)	Movie Subtitles	8+neu	Yes	53.6%	54.4%	Y

In this paper, we consider several variants of the F1 score as well as precision-recall curves. The outcomes of our analysis are presented in the results section.

3. Results

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 by the use of 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.

As discussed earlier, the micro F1 score is expected to perform better when classes are balanced, as it aggregates contributions across all classes, without considering their labels. Conversely, the macro F1 score is preferred over the micro F1 in the presence of imbalanced classes, because it emphasizes performance on minority classes. The weighted F1 score is suited for scenarios with class imbalance, offering a measure of overall classifier performance that accounts for class distribution. Which F1 score is preferred over the other, depends on what is important to evaluate in a classifier: the performance of the minority class, or the performance considering the contribution of each class.

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 considered to be greater than 0.7 (Joshi et al., 2022). 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 (Coden et al., 2009).

In emotion analysis, studies have achieved macro F1 scores of 0.53, 0.62, and 0.64 (Öhman et al., 2020; Kane et al., 2022; Demszky et al., 2020) using techniques such as BERT, which employs transformer-based models pretrained and fine-tuned to achieve state-of-the-art results (Devlin et al., 2018). Comparing against a baseline model helps determine what constitutes an acceptable F1 score (Rajpurkar et al., 2016). For this experiment, the F1 score —micro, macro, or weighted—of 0.53 achieved on the XED dataset will serve as our benchmark for a good F1 score. It's important to note that while BERT considers contextual information, the lexicons used in our experiment only consist of unigrams (Devlin et al., 2018; Öhman et al., 2020).

Each distilled lexicon was applied to the datasets to classify per emotion the messages of the datasets. Afterwards, the F1 scores were calculated for each emotion category, with thresholds specific to each lexicon. Precision-recall (PR) curves were generated for each emotion category, reflecting the varying thresholds to generate the lexicons (Buckland & Gey, 1994; Davis & Goadrich, 2006). An ideal PR curve achieves both precision and recall scores of one, with an area under the curve (AUC) ideally equal to 1. These curves illustrate the trade-off between

precision and recall, with curves closer to the top right corner indicating superior model performance (Buckland & Gey, 1994; Davis & Goadrich, 2006). Figure 4 illustrates examples of different precision-recall curves. The best curve is the one closer the top right corner, or the one with the highest the area under the curve.

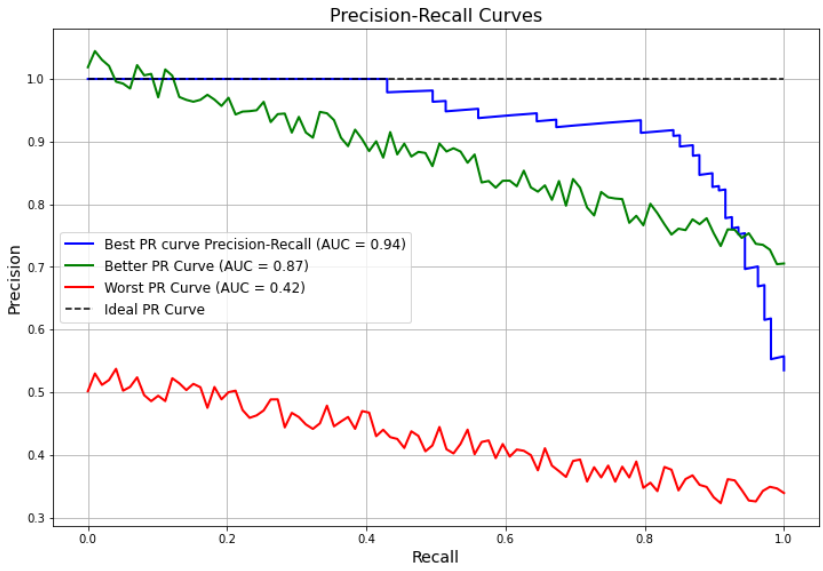


FIGURE 4: EXAMPLES OF PRECISION-RECALL (PR) CURVES

3.1 Overview of Experiments

In this study, several experiments were conducted to evaluate the performance of different F1 score variants—micro, macro, and weighted—using lexicon-based emotion analysis models. The key experiments involve applying derived lexicons at different thresholds, to various datasets. The experiments are summarized as follows:

1. GoEmotions Lexicon Applied to GoEmotions Dataset:
    - A lexicon was derived at different thresholds from the GoEmotions dataset and applied back to the same dataset to evaluate its performance across six different emotions.
2. GoEmotions Lexicon Applied to XED Dataset:
    - The GoEmotions-derived lexicon at different thresholds was applied to the XED dataset to assess its generalizability and performance on a different dataset.
3. XED Lexicon Applied to XED Dataset:
    - The Lexicon was distilled from the XED dataset at different thresholds and applied to the same dataset to evaluated its performance across the same six emotions.
4. XED Lexicon Applied to GoEmotions Dataset:
    - The XED-derived lexicon was applied to the GoEmotions dataset to test its applicability and performance on a different dataset.

5. GoEmotions Aggregated Lexicon Applied to GoEmotions Aggregated Dataset:
    - As explained in section 2.1, for the aggregated datasets the emotions anger, disgust, fear, and sadness were consolidated into the category of "negative". This dataset has 3 categories: 'surprise,' 'joy,' and 'negative'. The lexicon derived from this aggregated dataset at different thresholds, was applied back to the same aggregated dataset.
6. GoEmotions Aggregated Lexicon Applied to XED Aggregated Dataset:
    - The aggregated lexicon derived from the GoEmotions dataset at different thresholds, was applied to the aggregated XED dataset, where the emotions were also grouped in the same way.
7. XED Aggregated Lexicon Applied to XED Aggregated Dataset:
    - A lexicon was derived from the aggregated XED dataset and evaluated its performance on the same aggregated dataset.
8. XED Aggregated Lexicon Applied to GoEmotions Aggregated Dataset:
    - The aggregated lexicon derived from the XED dataset was applied to the aggregated GoEmotions dataset.

3.2 Precision-Recall Curves

For each experiment, precision-recall (PR) curves were generated each of the six emotions from Ekman or the three aggregated emotion categories. The points on each curve represent different thresholds used in the evaluation.

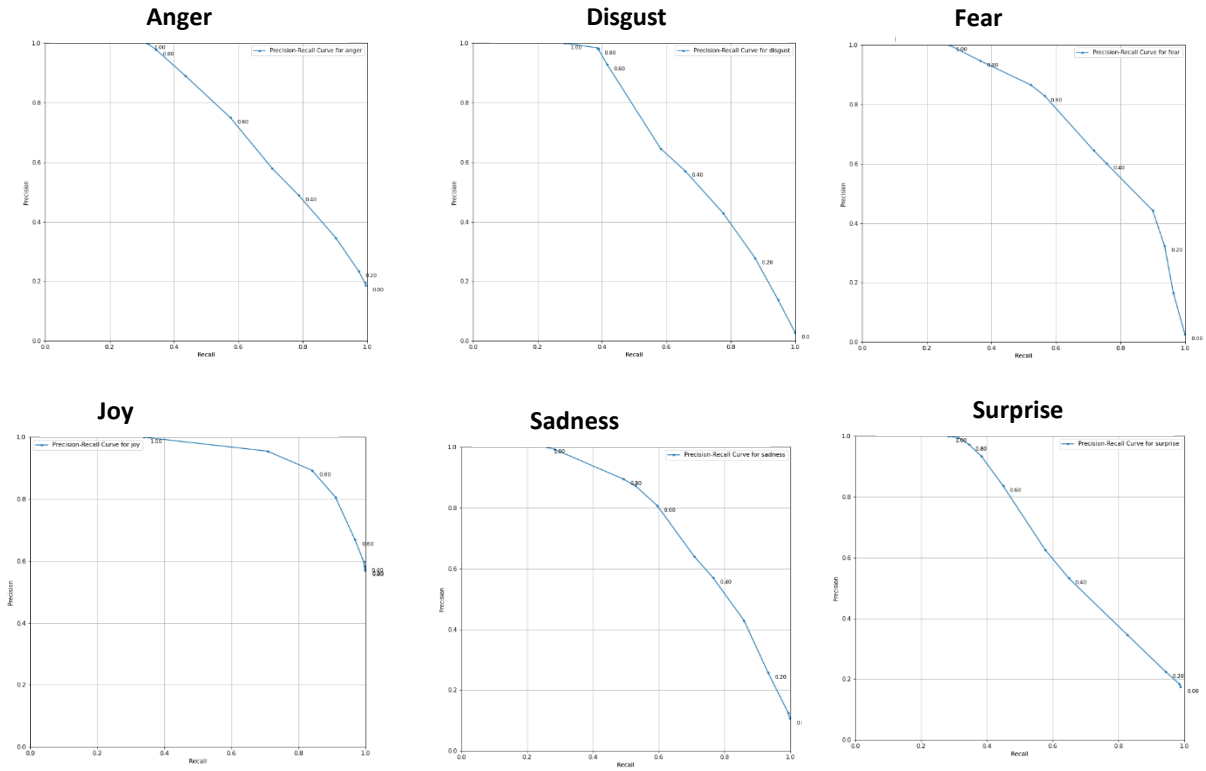


FIGURE 5: PRECISION-RECALL CURVES FOR SIX DIFFERENT EMOTIONS ANALYZED USING THE GOEMOTIONS DISTILLED LEXICON AND APPLIED TO THE GOEMOTIONS DATASET. EACH GRAPHIC REPRESENTS THE PRECISION-RECALL CURVE FOR A SPECIFIC EMOTION, WITH POINTS ON THE CURVE REPRESENTING DIFFERENT THRESHOLDS.

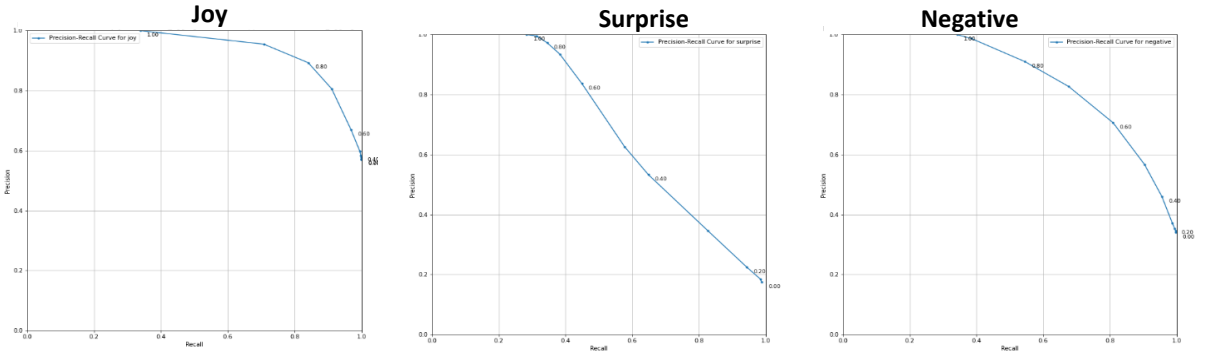


FIGURE 6: PRECISION-RECALL CURVES FOR THREE DIFFERENT EMOTIONS ANALYZED USING THE GOEMOTIONS AGGREGATED DISTILLED LEXICON AND APPLIED TO THE GOEMOTIONS AGGREGATED DATASET. EACH GRAPHIC REPRESENTS THE PRECISION-RECALL CURVE FOR A SPECIFIC EMOTION, WITH POINTS ON THE CURVE REPRESENTING DIFFERENT THRESHOLDS.

Figure 5 displays the precision-recall curves of all 6 emotions of the GoEmotions distilled lexicon, applied to the GoEmotions dataset. In the case of the emotion "Joy", the curve demonstrates high recall and precision, indicating robust model performance for this emotion. Similarly, Figure 6 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.

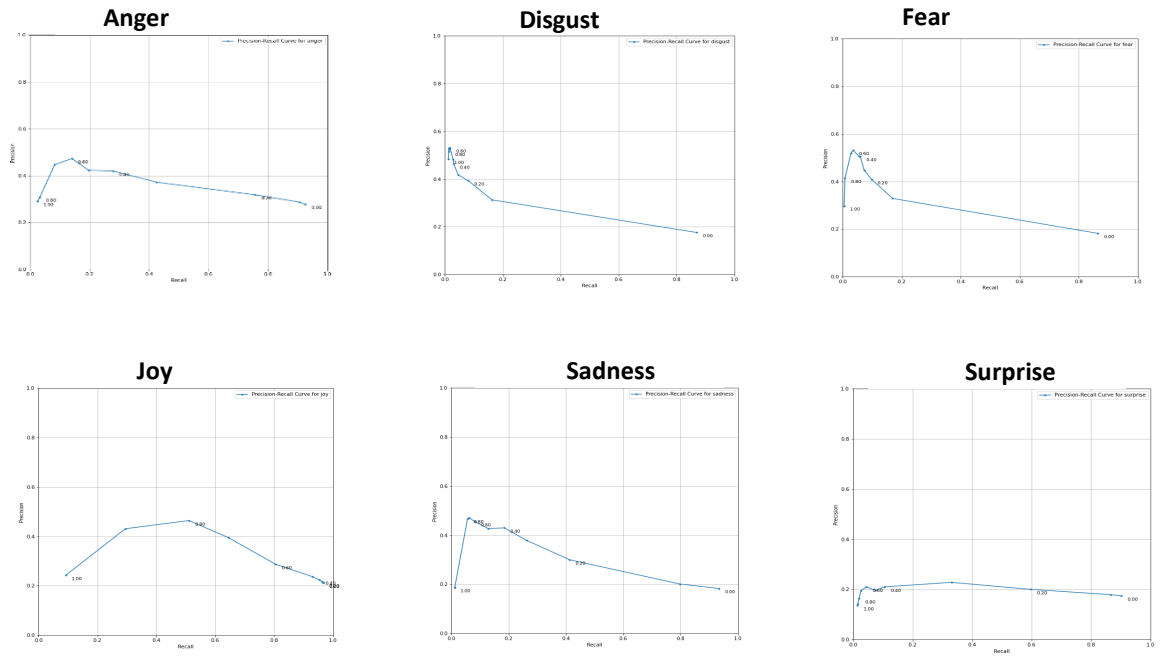


FIGURE 7: PRECISION-RECALL CURVES FOR SIX DIFFERENT EMOTIONS ANALYZED USING THE GOEMOTIONS DISTILLED LEXICON AND APPLIED TO THE XED DATASET. EACH GRAPHIC REPRESENTS THE PRECISION-RECALL CURVE FOR A SPECIFIC EMOTION, WITH POINTS ON THE CURVE REPRESENTING DIFFERENT THRESHOLDS.

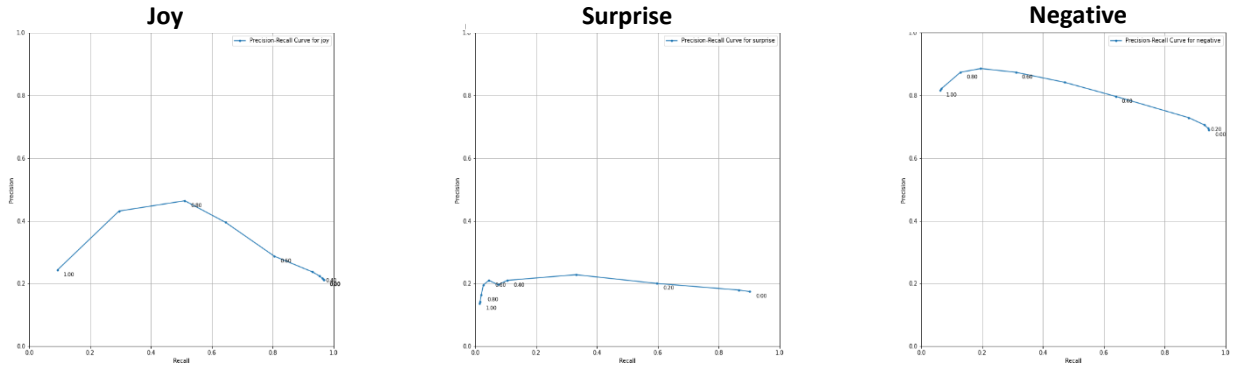


FIGURE 8: PRECISION-RECALL CURVES FOR THREE DIFFERENT EMOTIONS ANALYZED USING THE GOEMOTIONS AGGREGATED DISTILLED LEXICON AND APPLIED TO THE XED AGGREGATED DATASET. EACH GRAPHIC REPRESENTS THE PRECISION-RECALL CURVE FOR A SPECIFIC EMOTION, WITH POINTS ON THE CURVE REPRESENTING DIFFERENT THRESHOLDS.

Figure 7 depicts the precision-recall curves for the GoEmotions distilled lexicon applied to the XED dataset. This visualization reveals a notably low recall and precision across the curves, indicating that the model is not very effective for this particular application.

Conversely, Figure 8 illustrates the precision-recall curve for 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.

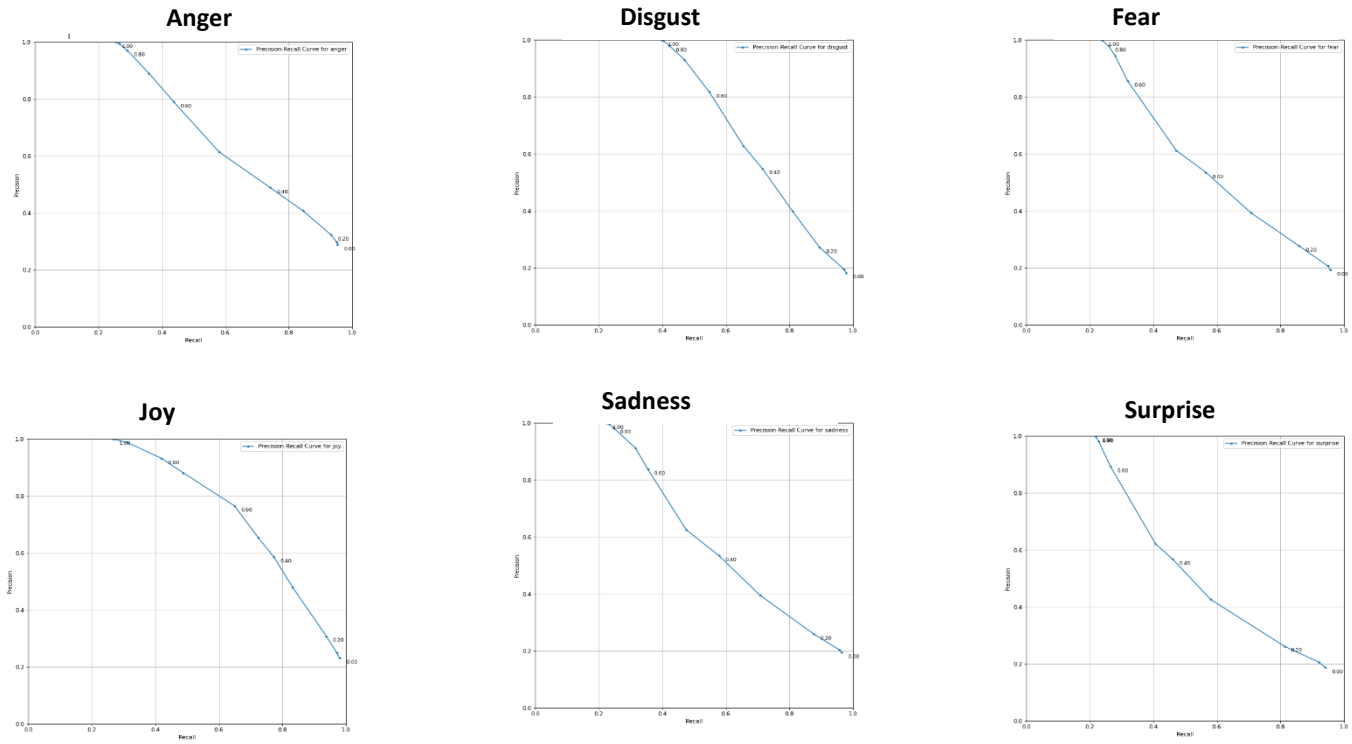


FIGURE 9: PRECISION-RECALL CURVES FOR SIX DIFFERENT EMOTIONS ANALYZED USING THE XED DISTILLED LEXICON AND APPLIED TO THE XED DATASET. EACH GRAPHIC REPRESENTS THE PRECISION-RECALL CURVE FOR A SPECIFIC EMOTION, WITH POINTS ON THE CURVE REPRESENTING DIFFERENT THRESHOLDS.

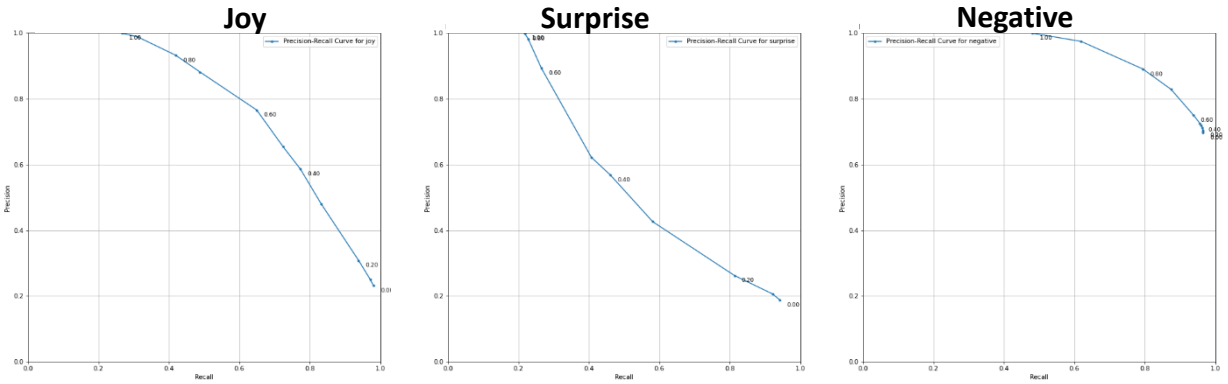


FIGURE 10: PRECISION-RECALL CURVES FOR THREE DIFFERENT EMOTIONS ANALYZED USING THE XED AGGREGATED DISTILLED LEXICON AND APPLIED TO THE XED AGGREGATED DATASET. EACH GRAPHIC REPRESENTS THE PRECISION-RECALL CURVE FOR A SPECIFIC EMOTION, WITH POINTS ON THE CURVE REPRESENTING DIFFERENT THRESHOLDS.

Figure 9 displays the precision-recall curves for the XED distilled lexicon applied to the XED dataset. The curves demonstrate high recall and precision for 'joy' and the other emotions, indicating robust model performance across different emotional categories. The precision-recall curves for the XED aggregated distilled lexicon and applied to the XED aggregated dataset (Figure 10) also shows an improvement in the precision-recall curve in the negative emotions, compared to the non-aggregated curves.

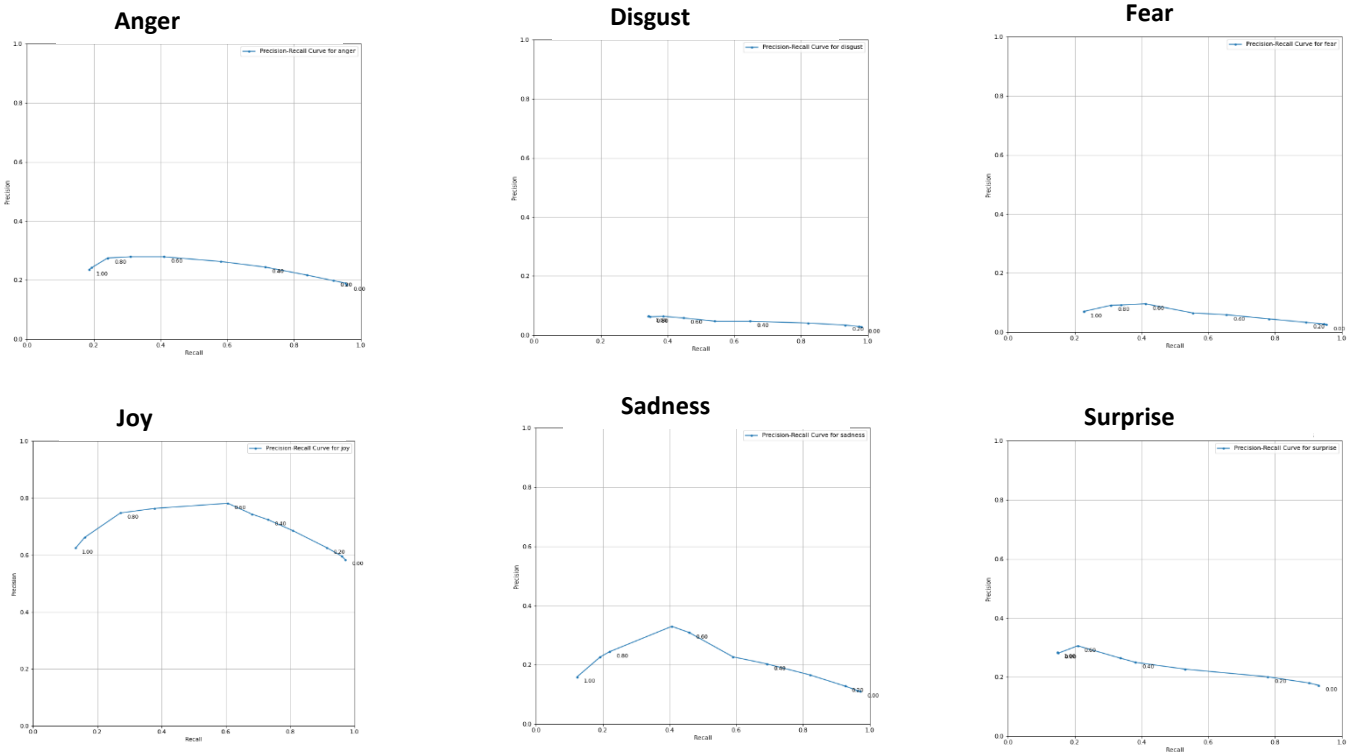


FIGURE 11: PRECISION-RECALL CURVES FOR SIX DIFFERENT EMOTIONS ANALYZED USING THE XED DISTILLED LEXICON AND APPLIED TO THE GOEMOTIONS DATASET. EACH GRAPHIC REPRESENTS THE PRECISION-RECALL CURVE FOR A SPECIFIC EMOTION, WITH POINTS ON THE CURVE REPRESENTING DIFFERENT THRESHOLDS.

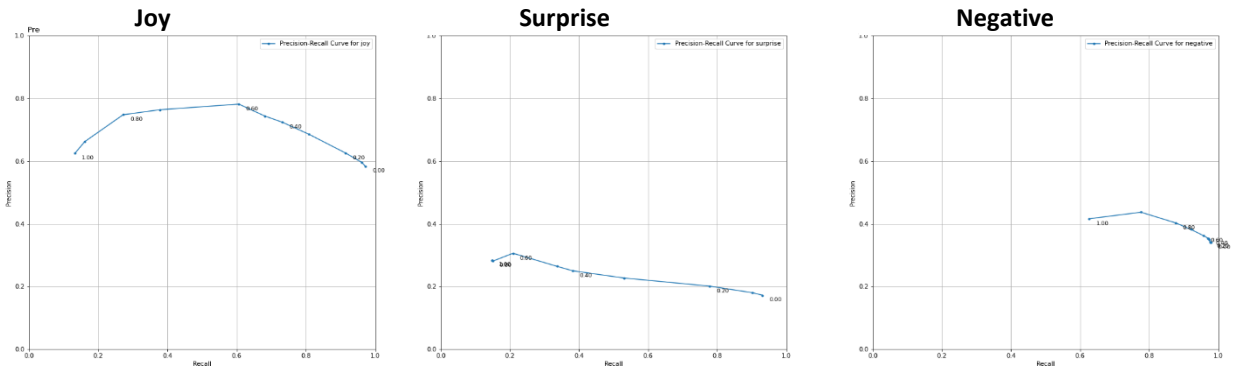


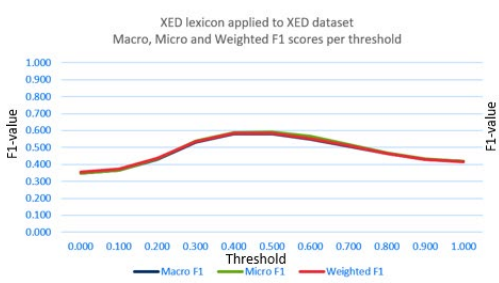
FIGURE 12: PRECISION-RECALL CURVES FOR THREE DIFFERENT EMOTIONS ANALYZED USING THE XED AGGREGATED DISTILLED LEXICON AND APPLIED TO THE GOEMOTIONS AGGREGATED DATASET. EACH GRAPHIC REPRESENTS THE PRECISION-RECALL CURVE FOR A SPECIFIC EMOTION, WITH POINTS ON THE CURVE REPRESENTING DIFFERENT THRESHOLDS.

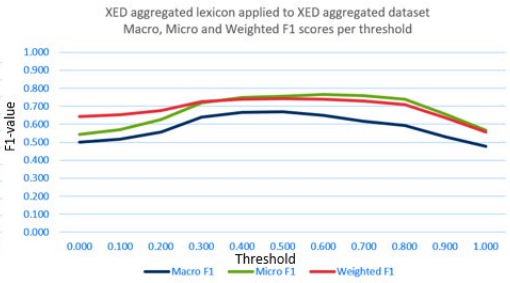
Figure 11 illustrates the precision-recall curves for the XED distilled lexicon applied to the GoEmotions dataset. These curves indicate high precision and recall for the emotion 'joy,' suggesting effective model performance for this specific emotion. Similarly, the performance metrics for aggregated negative emotions, shown in Figure 12, also show improvement. These precision-recall curves also demonstrate that the emotion 'joy' tends to be more accurately identified by the models. A plausible explanation for this observation 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 categories, unlike the negative emotions, which might share overlapping expressive features, potentially confusing the model.

3.3 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.





Metrics XED lexicon applied to XED dataset

Threshold	Macro F1	Micro F1	Weighted F1
0.000	0.347	0.349	0.356
0.100	0.365	0.367	0.373
0.200	0.430	0.434	0.435
0.300	0.533	0.537	0.536
0.400	0.582	0.588	0.584
0.500	0.582	0.591	0.585
0.600	0.549	0.565	0.553
0.700	0.506	0.516	0.509
0.800	0.463	0.469	0.464
0.900	0.431	0.433	0.430
1.000	0.417	0.418	0.415

Metrics XED aggregated lexicon applied to XED aggregated dataset

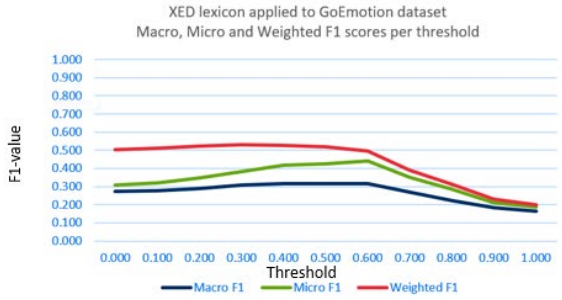
Threshold	Macro F1	Micro F1	Weighted F1
0.000	0.499	0.545	0.643
0.100	0.515	0.568	0.653
0.200	0.558	0.627	0.677
0.300	0.640	0.720	0.724
0.400	0.666	0.749	0.740
0.500	0.668	0.757	0.743
0.600	0.648	0.765	0.738
0.700	0.616	0.760	0.728
0.800	0.593	0.739	0.710
0.900	0.530	0.656	0.636
1.000	0.476	0.567	0.557

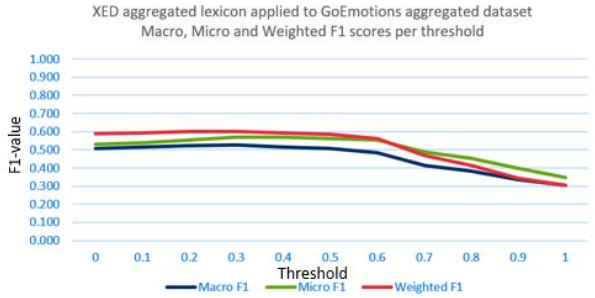
FIGURE 13: COMPARATIVE ANALYSIS OF THE XED DISTILLED LEXICON PERFORMANCE ON THE XED DATASET

ABOVE: CHART THAT DEPICTS MICRO, MACRO, AND WEIGHTED F1 SCORES FOR THE XED STANDARD AND AGGREGATED LEXICONS APPLIED TO THE XED STANDARD AND AGGREGATED DATASETS ACROSS VARIOUS THRESHOLDS.

BELOW: CORRESPONDING TABLES OF METRICS FOR THE XED STANDARD AND AGGREGATED LEXICONS, APPLIED TO THEIR RESPECTIVE DATASETS.

Figure 13 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.





Metrics XED lexicon applied to GoEmotions dataset

Threshold	Macro F1	Micro F1	Weighted F1
0.000	0.271	0.310	0.504
0.100	0.277	0.320	0.510
0.200	0.290	0.346	0.522
0.300	0.308	0.384	0.530
0.400	0.317	0.416	0.528
0.500	0.317	0.426	0.519
0.600	0.315	0.443	0.497
0.700	0.269	0.353	0.389
0.800	0.221	0.284	0.313
0.900	0.182	0.212	0.229
1.000	0.163	0.188	0.201

Metrics XED aggregated lexicon applied to GoEmotions aggregated dataset

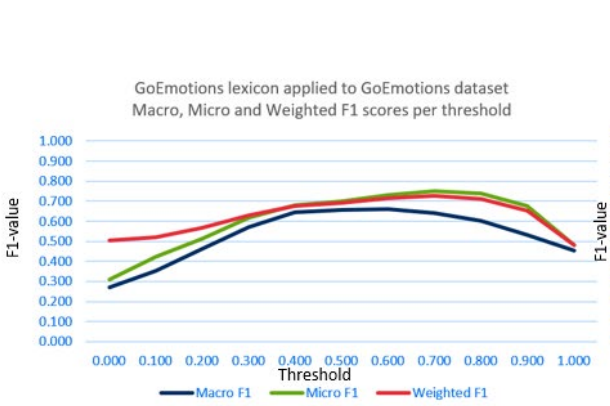
Threshold	Macro F1	Micro F1	Weighted F1
0	0.508	0.530	0.588
0.1	0.514	0.538	0.593
0.2	0.524	0.556	0.601
0.3	0.525	0.569	0.603
0.4	0.516	0.569	0.594
0.5	0.509	0.563	0.584
0.6	0.486	0.556	0.563
0.7	0.414	0.487	0.467
0.8	0.382	0.451	0.414
0.9	0.338	0.397	0.342
1	0.305	0.348	0.303

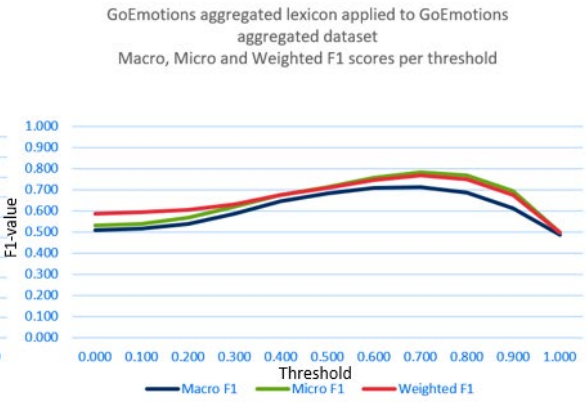
FIGURE 14:COMPARATIVE ANALYSIS OF THE XED DISTILLED LEXICON PERFORMANCE ON THE GOEMOTIONS DATASET

ABOVE: CHART THAT DEPICTS MICRO, MACRO, AND WEIGHTED F1 SCORES FOR THE XED STANDARD AND AGGREGATED LEXICONS APPLIED TO THE GOEMOTIONS STANDARD AND AGGREGATED DATASETS ACROSS VARIOUS THRESHOLDS.

BELOW: CORRESPONDING TABLES OF METRICS FOR THE XED STANDARD AND AGGREGATED LEXICONS, APPLIED TO THEIR RESPECTIVE DATASETS.

Figure 14 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 cut off of 0.53, which suggests suboptimal model performance . In contrast, the weighted F1 score indicates an acceptable performance level, in 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.





Metrics GoEmotions lexicon applied to GoEmotions dataset

Threshold	Macro F1	Micro F1	Weighted F1
0.000	0.272	0.310	0.503
0.100	0.351	0.421	0.520
0.200	0.462	0.514	0.565
0.300	0.573	0.618	0.630
0.400	0.643	0.681	0.677
0.500	0.658	0.700	0.692
0.600	0.660	0.732	0.716
0.700	0.642	0.752	0.728
0.800	0.601	0.738	0.710
0.900	0.532	0.675	0.652
1.000	0.452	0.480	0.480

GoEmotions aggregated lexicon applied to GoEmotions aggregated dataset

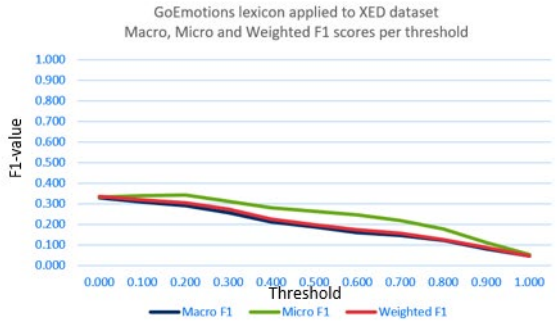
Threshold	Macro F1	Micro F1	Weighted F1
0.000	0.510	0.531	0.588
0.100	0.517	0.540	0.593
0.200	0.537	0.569	0.605
0.300	0.586	0.619	0.632
0.400	0.647	0.676	0.676
0.500	0.681	0.713	0.708
0.600	0.710	0.757	0.746
0.700	0.714	0.785	0.770
0.800	0.686	0.768	0.751
0.900	0.614	0.695	0.677
1.000	0.486	0.498	0.497

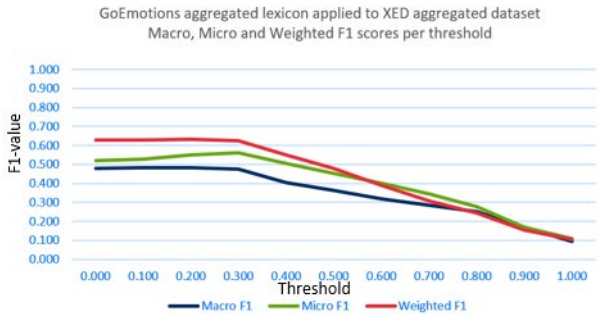
FIGURE 15: COMPARATIVE ANALYSIS OF THE GOEMOTIONS DISTILLED LEXICON PERFORMANCE ON THE GOEMOTIONS DATASET

ABOVE: CHART THAT DEPICTS MICRO, MACRO, AND WEIGHTED F1 SCORES FOR THE GOEMOTIONS STANDARD AND AGGREGATED LEXICONS APPLIED TO THE GOEMOTIONS STANDARD AND AGGREGATED DATASETS ACROSS VARIOUS THRESHOLDS.

BELOW: CORRESPONDING TABLES OF METRICS FOR THE GOEMOTIONS STANDARD AND AGGREGATED LEXICONS, APPLIED TO THEIR RESPECTIVE DATASETS.

Figure 15 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. This performance pattern is mirrored in the results from the aggregated GoEmotions lexicon applied to the aggregated GoEmotions dataset. 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.





Metrics GoEmotions lexicon applied to XED dataset

Threshold	Macro F1	Micro F1	Weighted F1
0.000	0.328	0.330	0.337
0.100	0.307	0.339	0.319
0.200	0.289	0.341	0.305
0.300	0.257	0.311	0.272
0.400	0.210	0.279	0.226
0.500	0.185	0.262	0.198
0.600	0.159	0.245	0.171
0.700	0.145	0.219	0.154
0.800	0.121	0.177	0.125
0.900	0.081	0.109	0.085
1.000	0.044	0.051	0.046

Metrics GoEmotions aggregated lexicon applied to XED aggregated dataset

Threshold	Macro F1	Micro F1	Weighted F1
0.000	0.478	0.520	0.626
0.100	0.481	0.526	0.629
0.200	0.484	0.549	0.632
0.300	0.474	0.560	0.624
0.400	0.404	0.505	0.548
0.500	0.363	0.453	0.478
0.600	0.319	0.401	0.388
0.700	0.286	0.344	0.308
0.800	0.249	0.278	0.244
0.900	0.167	0.170	0.151
1.000	0.093	0.107	0.106

FIGURE 16:COMPARATIVE ANALYSIS OF THE GOEMOTIONS DISTILLED LEXICON PERFORMANCE ON THE XED DATASET

ABOVE: CHART THAT DEPICTS MICRO, MACRO, AND WEIGHTED F1 SCORES FOR THE GOEMOTIONS STANDARD AND AGGREGATED LEXICONS APPLIED TO THE XED STANDARD AND AGGREGATED DATASETS ACROSS VARIOUS THRESHOLDS.

BELOW: CORRESPONDING TABLES OF METRICS FOR THE GOEMOTIONS STANDARD AND AGGREGATED LEXICONS, APPLIED TO THEIR RESPECTIVE DATASETS.

Figure 16 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. This signifies a 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 cut off 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. In this case, the micro and weighted F1 scores, which are higher and surpass the 0.53 threshold.

The behavior of the micro F1 score, as observed in Figure 13 and Figure 16, 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 to aggregate 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.

The weighted F1 score demonstrates that it is more 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 it can be seen in Table 1, the XED dataset is balanced. This trend of 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 a better performance. It is also a metric that is responsive to changes in class balance and in this case, word availability.

4. Discussion

This study explored the performance of micro, macro, and weighted F1 scores using lexicon-based emotion analysis models applied to the GoEmotions and XED datasets. The experiments aimed to understand how these metrics perform in the presence of class imbalance—a common challenge in emotion analysis. Several key insights were revealed and are discussed below.

Regarding the micro F1 score, this metric performed as well as the other metrics in balanced datasets because it aggregates contributions across all classes. When applying the GoEmotions distilled lexicon to the balanced XED dataset, the micro F1 score showed higher values compared to other F1 variants.

In contrast, the macro F1 score highlighted performance disparities in imbalanced datasets. This evaluation metric penalizes models that underperform on minority classes. This was evident when applying the XED lexicon to the GoEmotions dataset. The macro F1 score's sensitivity to the performance of minority classes is key for situations that require equitable performance across all classes.

The weighted F1 score provided a view of the overall classifier performance by considering class distribution. This suggests that the weighted F1 score is useful for scenarios with class imbalance, as it offers a perspective that considers the prevalence of each class. This can be preferable in situations where the minority class is not expected to perform as well as the majority class. In the experiments, when using the aggregated datasets, the weighted F1 score had higher values compared to the macro and micro F1 scores.

Another significant finding was the impact of the dataset configuration. Aggregating emotions into fewer categories improved model performance metrics, particularly for macro F1 scores. This suggests that class aggregation can mitigate the effects of class imbalance. This issue also highlights a limitation of using the Ekman emotion classification system, which considers only one positive emotion and four negative ones. This can introduce class imbalance and affect the classifier performance on minority classes.

The main limitation of this study is that only the Ekman emotion classification system was used. Furthermore, only unigrams were considered by using the distilled lexicons, which meant that context was not taken into account, which impacted the performance. Future research should explore the influence of class imbalance and category aggregation on emotion classification models, considering different emotion systems and the effect that aggregating emotions has in those cases.

5. Conclusion

This study confirms the importance of selecting appropriate evaluation metrics in emotion classification tasks, particularly in the context of class imbalance. Specifically, it points out that the use of only one metric does not give a nuanced image of the classifier performance.

The micro F1 score was effective at providing a general view of the performance of the model in balanced datasets because every class is contributing evenly to the result.

The macro F1 score is valuable for identifying performance discrepancies in imbalanced datasets, as it penalized an underperformance of the classifier in the minority class. Depending on the research objective, this bias may not be desirable. For example, in multilabel emotion analysis, this penalization might not be an advantage when the researcher is interested in emotions that are the majority classes, and there is no interest in the minority class. Using only the macro F1 score as a metric might give a false picture of the usefulness of a dataset or a model.

The weighted F1 score offers a comprehensive view of model performance that accounts for class distribution, making it useful for imbalanced scenarios. In this case, it can help to give an image of the overall performance of the classification model.

This research also highlights the benefit of aggregating emotions to improve performance metrics. We propose that further research should be focused on comparing diverse emotion systems to understand their influence on class imbalance.

In conclusion, presenting the weighted, macro, and micro F1 scores offers a more comprehensive evaluation of the performance of a classifier in imbalance scenarios than only selecting one metric. This balanced approach for evaluating classifiers using imbalanced multilabel datasets can provide valuable insights into emotion classification tasks, when comparing models. We think that these metrics should be used together with the F1 scores per class for multilabel classification tasks. This could avoid the issues caused by averaging them and assigning weights.

6. Acknowledgments

For this paper, ChatGPT and Grammarly were used to provide assistance in refining the grammar, and improving the flow and clarity of the text. This was beneficial, as English is not the author's native language.

7. References

Abbasi, B., & Goldenholz, D. M. (2019). Machine learning applications in epilepsy. Epilepsia, 60(10), 2037–2047. https://doi.org/10.1111/epi.16333

Abdul-Mageed, M., & Ungar, L. H. (2017). EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks. Association for Computational Linguistics. https://doi.org/10.18653/v1/p17-1067

Akosa, J. S. (2017). Predictive accuracy: a misleading performance measure for highly imbalanced data.

Alpaydin, E. (2016). Machine learning: The New AI. MIT Press.

Andrikakis, E., Perikos, I., Paraskevas, M., & Hatzilygeroudis, I. (2023). Text analysis and recognition of emotional content using deep learning methods and BERT. https://doi.org/10.1109/icis57766.2023.10210232

Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. Journal of Information Engineering and Applications, 3(10), 27–38. https://www.iiste.org/Journals/index.php/JIEA/article/download/7633/8051

Buckland, M. & Gey, F. (1994). The relationship between recall and precision. Journal of the American Society for Information Science and Technology. https://doi.org/10.5555/184656.180369

Buttar, J. K. P. K. (2018, April 30). A Systematic Review on StopWord Removal Algorithms. http://www.ijfrcsce.org/index.php/ijfrcsce/article/view/1499

Coden, A., Savova, G., Sominsky, I., Tanenblatt, M., Masanz, J., Schuler, K., Cooper, J., Guan, W., & De Groen, P. C. (2009). Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. Journal of Biomedical Informatics, 42(5), 937–949. https://doi.org/10.1016/j.jbi.2008.12.005

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1). https://doi.org/10.1186/s12864-019-6413-7

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning. https://doi.org/10.1145/1143844.1143874

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A Dataset of Fine-Grained Emotions. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2005.00547

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1810.04805

Ekman, P. (1992). An argument for basic emotions. Cognition and Emotion, 6(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Erickson, B. J., & Kitamura, F. (2021). Magician's corner: 9. Performance Metrics for Machine learning models. Radiology. Artificial Intelligence, 3(3), e200126. https://doi.org/10.1148/ryai.2021200126

Fang, J. (2023). The role of data imbalance bias in the prediction of protein stability change upon mutation. PloS One, 18(3), e0283727. https://doi.org/10.1371/journal.pone.0283727

Flach, P. A., & Kull, M. (2015). Precision-Recall-Gain Curves: PR Analysis done right. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), Advances in Neural Information Processing Systems 28 (Vols. 1–1, pp. 838–846). Massachusetts Institute of Technology (MIT) Press. https://research-information.bris.ac.uk/ws/portalfiles/portal/72164009/5867_precision_recall_gain_curves_pr_analysis_done_right.pdf

Ford, V. & Ambareen Siraj. (2014). Applications of machine learning in cyber Security. In Conference Paper [Conference-proceeding]. https://www.researchgate.net/publication/283083699

Gu, Q., Zhu, L., & Cai, Z. (2009). Evaluation measures of the classification performance of imbalanced data sets. In Communications in computer and information science (pp. 461–471). https://doi.org/10.1007/978-3-642-04962-0_53

Gupta, A., Anand, A., & Hasija, Y. (2021). Recall-based Machine Learning approach for early detection of Cervical Cancer. IEEE. https://doi.org/10.1109/i2ct51068.2021.9418099

Harbecke, D., Chen, Y., Hennig, L., & Alt, C. (2022). Why only Micro-F1? Class Weighting of Measures for Relation Classification. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2205.09460

Honnibal, M., & Montani, I. (2023). SpaCy (Version 3.7.1) [Software]. Explosion AI. Available from https://spacy.io

Huang, C., Trabelsi, A., Qin, X., Farruque, N., & Zäıane, O. R. (2019). SEQ2EMO for multi-label emotion classification based on latent variable chains transformation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1911.02147

Hugging Face. (2023). GoEmotions and XED datasets. Hugging Face, Inc. Retrieved from https://huggingface.co/datasets

Joshi, K., Kumar, S., Rawat, J., Kumari, A., Gupta, A., & Sharma, N. (2022). Fraud app detection of Google Play Store apps using Decision tree. 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM). https://doi.org/10.1109/iciptm54933.2022.9754207

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1). https://doi.org/10.1186/s40537-019-0192-5

Kane, A., Patankar, S., Khose, S., & Kirtane, N. (2022). Transformer based ensemble for emotion detection. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2203.11899

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2020). Fundamentals of Machine Learning for Predictive Data Analytics, second edition: Algorithms, Worked Examples, and Case Studies. MIT Press.

Lango, M. (2019). Tackling the problem of class imbalance in multi-class sentiment Classification: an experimental study. Foundations of Computing and Decision Sciences, 44(2), 151–178. https://doi.org/10.2478/fcds-2019-0009

Liu, C., Osama, M., & De Andrade, A. (2019). DENS: A Dataset for Multi-class Emotion Analysis. Association for Computational Linguistics. https://doi.org/10.18653/v1/d19-1656

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2019). Overview of the HASOC track at FIRE 2019. Association for Computing Machinery. https://doi.org/10.1145/3368567.3368584

Miao, J., & Zhu, W. (2021). Precision–recall curve (PRC) classification trees. Evolutionary Intelligence, 15(3), 1545–1569. https://doi.org/10.1007/s12065-021-00565-2

Öhman, E., Pàmies, M., Kajava, K., & Tiedemann, J. (2020). XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection. International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.575

Olek, M. (2023). About evaluation of F1 score for RECENT Relation Extraction system. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2305.09410

Poudel, S. (2022). A study of disease diagnosis using Machine Learning. Medical Sciences Forum. https://doi.org/10.3390/iech2022-12311

Powers, D. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2010.16061

Rainio, O., Teuho, J., & Клен, P. (2024). Evaluation metrics and statistical tests for machine learning. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-56706-x

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. The New England Journal of Medicine, 1347–1358. https://doi.org/10.1056/NEJMra1814259

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. Association for Computational Linguistics. https://doi.org/10.18653/v1/d16-1264

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PloS One, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Strapparava, C., & Mihalcea, R. (2007, June 1). SEmEVAL-2007 Task 14: Affective Text. ACL Anthology. https://aclanthology.org/S07-1013

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). CLASSIFICATION OF IMBALANCED DATA: a REVIEW. International Journal of Pattern Recognition and Artificial Intelligence, 23(04), 687–719. https://doi.org/10.1142/s0218001409007326

Takahashi, K., Yamamoto, K., Kuchiba, A., & Koyama, T. (2021). Confidence interval for micro-averaged F1 and macro-averaged F1 scores. *Applied Intelligence*, *52*(5), 4961–4972. https://doi.org/10.1007/s10489-021-02635-5

Ting, K.M. (2011). Precision and Recall. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_652

Van Rijsbergen, C. J. (1979). Information retrieval. London; Toronto: Butterworths.

Wang, H., Xu, Q., & Zhou, L. (2015). Large unbalanced credit scoring using Lasso-Logistic Regression ensemble. PloS One, 10(2), e0117844. https://doi.org/10.1371/journal.pone.0117844

Xu, R., Chen, T., Xia, Y., Lu, Q., Liu, B., & Wang, X. (2015). Word embedding composition for data imbalances in sentiment and emotion classification. Cognitive Computation, 7(2), 226–240. https://doi.org/10.1007/s12559-015-9319-y