Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1392–1399
Marseille, 20-25 June 2022

RED v2: Enhancing RED Dataset for Multi-Label Emotion Detection

Alexandra Ciobotaru[1], **Mihai V. Constantinescu**[2]
Liviu P. Dinu[1], **Stefan Daniel Dumitrescu**[3]
University of Bucharest, Faculty of Mathematics and Computer Science[1],
Independent Researcher[2] Adobe[3]
Bucharest, Romania
alexandra.ciobotaru@unibuc.ro, mihai.vlad6@gmail.com
ldinu@fmi.unibuc.ro, sdumitre@adobe.com

Abstract

RED (Romanian Emotion Dataset) is a machine learning-based resource developed for the automatic detection of emotions in Romanian texts, containing single-label annotated tweets with one of the following emotions: joy, fear, sadness, anger and neutral. In this work, we propose *REDv2*, an open-source extension of RED by adding two more emotions, trust and surprise, and by widening the annotation schema so that the resulted novel dataset is multi-label. We show the overall reliability of our dataset by computing inter-annotator agreements per tweet using a formula suitable for our annotation setup and we aggregate all annotators' opinions into two variants of ground truth, one suitable for multi-label classification and the other suitable for text regression. We propose strong baselines with two transformer models, the Romanian BERT and the multilingual XLM-Roberta model, in both categorical and regression settings.

Keywords: emotion detection, multi-label classification, text regression, Romanian tweets

1. Introduction

Interpreting correctly one's own emotions, as well as other people's emotional states, is a central aspect of emotional intelligence. Today, people can automate the process of emotion detection by creating machine learning models, provided by the fact that the model training was done on qualitative and sufficient data. With the constant increase of social media usage there is also an increase in online public data, freely available for model creation. Thus, analyzing emotions in online content naturally has became more and more of a topic of interest in the recent years.

Sentiment analysis, along with emotion analysis, are key technologies used to gain insights from social media networks, the opinion mining field reaching a level of maturity, sufficient enough to be used in practical applications (Iglesias and Moreno, 2019).

There are many definitions to emotions (Peng et al., 2021), but as defined in (Liu, 2015), an emotion is "a mental state that arises spontaneously rather than through a conscious effort; it is also often accompanied by physiological changes.". Many psychologists tried to create a taxonomy of emotions, but probably one of the most agreed upon taxonomy is Robert Plutchick's wheel of emotions (Plutchik, 1973). He identified eight primary emotions: happiness, trust, fear, surprise, sadness, disgust, anger and anticipation.

Having a model that automatically detects emotions in text has a wide range of applications, from computing the overall opinion of clients and/or potential customers in the field of brand management (Istrati and Ciobotaru, 2022), to automatic adaptation of chatbot answers in respect to the user's emotional state.

The first dataset of single labeled texts for detecting emotions from Romanian content is REDv1 (Romanian Emotion Dataset) (Ciobotaru and Dinu, 2021), a dataset containing roughly 4000 tweets annotated for the following emotions: fear, anger, happiness, sadness and neutral. Starting from this work, we expand REv1 by adding two more classes of emotions, surprise and trust, and also by increasing the overall number of texts and by widening the annotation schema to multi-label. In Table 1 we present a sample of our novel dataset, along with English translations, to aid the non-Romanian readers of this article.
The main contributions of this work are:

1. Expanded REDv1 (Ciobotaru and Dinu, 2021) by increasing the size of the dataset, by thoroughly cleaning it (see inter-annotator considerations), and by expanding it from single to a multi-label annotation schema. Thus REDv2 represents the *largest Twitter emotion dataset for the Romanian language*.

2. We provide baselines and validate the dataset with strong transformer-based models.

3. The dataset, models and evaluation scripts are open-source, freely available at https://github.com/Alegzandra/RED-Romanian-Emotions-Dataset/tree/main/REDv2

2. Recent Works

There are mainly two approaches that tackle the problem of emotion detection - the lexicon based approach and the machine learning approach (Canales

Text	Emotion(s)
Ca orice lucru nasol, incepe lunea	Tristețe
Like every bad thing, it starts monday	*Sadness*
Mulțumim frumos, sunt mândră de tine! Și noi vă iubim	Încredere, Bucurie
Thank you very much, I am proud of you! We love you too	*Trust, Happiness*
◁PROPN▷, șocată de cazul de dopaj de la ◁PROPN▷	Surpriză
◁PROPN▷, shocked about the doping case at ◁PROPN▷	*Surprise*

Table 1: Sample annotated texts from our dataset, with English translations

and Martínez-Barco, 2014). In this work we use a hybrid approach (Gievska et al., 2015) - we scrap tweets based on lexicon, manually verify them and based on those we create a multi-label dataset suitable for machine learning.

Detecting emotions from text is not a new research endeavour. One of the first created benchmarks for emotion detection in texts is Affect Text (Strapparava and Mihalcea, 2007), a dataset consisting of 1250 news headlines annotated for Anger, Disgust, Fear, Happiness, Sadness and Surprise created for SemEval-2007, task 14.

In English, there exist many datasets for emotion detection annotated using single labels, like ISEAR (Scherer and Wallbotth, 1994), WASSA (Mohammad and Bravo-Marquez, 2017) (this dataset also takes into account emotion intensity), just to name a few.

But naturally a text can contain more that one emotion, as one can express his/hers ideas in many different ways, the problem on emotion detection (ED) can be extrapolated to creating ED models trained on multi-label ED datasets.

SemEval-2018 task 1 (Mohammad et al., 2018) targeted detecting emotions in tweets using a multi-label dataset: Affect in Tweets Dataset (Mohammad and Kiritchenko, 2018). On this dataset (Jabreel and Moreno, 2019) obtains state-of-the-art results by using a method to convert the multiple emotions into a binary classification problem.

(Huang et al., 2019) create a balanced multi-label dataset containing emotional tweets, BMET (Balanced Multi-Label Emotional Tweets) and propose a novel model to solve the multi-label classification problem called Seq2Emo, a neural network that takes into account correlations between labels.

But probably the largest manually annotated dataset for multi-label emotion detection texts is GoEmotions (Demszky et al., 2020), containing 58k English Reddit comments, labeled for one or more of the 27 emotions, or neutral.

Emotion detection from Romanian texts is a domain researched previously by (Briciu and Lupea, 2017). They created the first lexicon for Romanian emotion detection, and further expanded the lexicon through formal concept analysis (Lupea and Briciu, 2019) where they create comparisons with another important lexical resource for Romanian, RoWordNet (Dumitrescu et al., 2018).

(Dinu et al., 2021) release a lexicon containing 770 Romanian pejorative words among three other pejorative lexicons for English, Spanish and Italian. Further, they release two datasets of annotated tweets containing pejorative words in English and Spanish tweets.

LaRoSeDa (A Large Romanian Sentiment Data Set) (Tache et al., 2021) is another important resource tangent to the field of emotion detection, comprised of 15,000 reviews in Romanian, annotated into positive and negative. Regarding sentiment analysis, (Istrati and Ciobotaru, 2022) explain in detail how they created a dataset of annotated Romanian tweets into positive and negative and create a baseline for sentiment detection in Romanian.

As far as we know, besides REDv1 (Ciobotaru and Dinu, 2021) there aren't any datasets for emotion detection in Romanian content. In this work we have improved REDv1 and we present it at its second version, with an enhanced number of labelled texts, two additional emotions and suitable for multi-label classification.

3. Dataset

Our dataset consists of 5449 tweets, labelled for one or more of the following emotions: sadness, surprise, fear, anger, trust, happiness, or neutral.

3.1. Scraping Process

Starting with (Ciobotaru and Dinu, 2021), we considered the work of (Mohammad and Bravo-Marquez, 2017) when creating the first annotated dataset in Romanian for detecting *fear*, *anger*, *happiness* and *sadness* in short texts, and added a *neutral* class, as it has been previously shown by (Al-Rubaiee et al., 2016) the importance of having a neutral class when classifying sentiments or emotions.

In this work, we took into consideration two more classes of emotions: *trust* and *surprise*, bringing the total number of labels per tweet to 7. Looking at the original (Ciobotaru and Dinu, 2021), the authors created lists of *query words* correspondent to each of the following classes: fear, anger, happiness, sadness and neutral. They scrapped tweets based on these words, manually annotated them (one label per tweet), leading to the creation of the first annotated dataset for emotion detection in Romanian short texts.

Class name	REDv1 QW	REDv2 QW
Anger	35	45
Fear	25	43
Happiness	32	39
Sadness	29	43
Surprise	0	28
Trust	0	26
Neutral	24	34

Table 2: Number of query words per class in REDv1 and in REDv2

Class name	Number of tweets
Anger	1336
Fear	1406
Happiness	1186
Sadness	1299
Surprise	726
Trust	1145
Neutral	852

Table 3: Total number of tweets after the first annotation step

In this work, we added query words from RoEmoLex, a lexicon of words developed for emotion analysis of Romanian texts, by (Briciu and Lupea, 2017). In Table 2 we present the total number of query words used for scrapping tweets for our improved dataset, REDv2, versus the number of tweets for the initial dataset, REDv1.

We scrapped tweets using the extra query words in order to improve the dataset by increasing the number of tweets per class, and creating two more classes: trust and sadness. Scraping was done using Snscrape[1] python library, in the time-frame: 1st of February 2020 - 1st of February 2021. All tweets were checked for Romanian using langdetect[2] python library.

3.2. Annotation Methods

3.2.1. First Annotation Step

The first annotation process involved 11 annotators, psychology students whose primary language is Romanian. They checked the scrapped tweets for each extra query word and kept a maximum of 50 tweets per query word. After the checking process was done, a number of 3973 new annotated tweets resulted. While REDv1 dataset contained 4047 annotated tweets, we concatenated it with the newly annotated tweets and it resulted a new dataset, containing 7947 annotated tweets, with the labels: sadness, happiness, fear, anger, surprise, trust and neutral.

The final number of tweets after the first step of the annotation process is detailed in Table 3.

All tweets were gathered from public accounts and, in order to protect confidentiality of Twitter users, we removed usernames and proper nouns from the final public dataset.

3.2.2. Second Annotation Step

The second annotation process involved 66 annotators, psychology students whose primary language is Romanian, including the 11 annotators from the first annotation step. They were each given sections consisting of 360 to 370 tweets. In order to facilitate the annotation process, we used the Doccano tool (Nakayama et al., 2018) where each annotator had access only to his dataset and every tweet could receive one or more labels or be marked as invalid. The annotators were instructed to select the Invalid label for the tweets they weren't sure about, or for the tweets that didn't make any more sense after removing usernames and proper nouns. Each of the 7947 unique tweets in the second batch was assigned to and annotated by 3 annotators.

3.3. Dataset preprocessing

From the resulted dataset we removed the tweets that received the Invalid label by at least one annotator, and also those tweets that had full mismatch between annotators, resulting a dataset containing 5449 tweets. Each annotation a tweet received was represented by a 3x7 matrix, which we will further call *annotation matrix*, with binary elements, 1 meaning that the text received a label for the emotion corresponding to the vector's index, and 0 otherwise. First line in matrix represents labels from Annotator 1, second line in matrix represents labels from Annotator 2 and the third line represents labels from Annotator 3. An exemplification of annotation matrix is shown in Figure 1.

['Sadness', 'Surprise', 'Fear', 'Anger', 'Neutral', 'Trust', 'Happiness']

$$\begin{matrix} \text{Ad1} \\ \text{Ad2} \\ \text{Ad3} \end{matrix} \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Figure 1: Annotation matrix with two labels: Surprise and Happiness.

In order to protect the anonymity of tweets authors as well as of the persons and institution discussed in texts and other sensitive data, we replaced the following text entities: proper nouns with ◁PROPN▷ (be it name or institution) using spacy ro_core_news_lg model[3] (Dumitrescu and Avram, 2019), email addresses with ◁EMAIL▷, telephone numbers with ◁TEL▷, usernames with ◁USERNAME▷ and links with ◁URL▷, using regex methods.

[1] https://github.com/JustAnotherArchivist/snscrape

[2] https://pypi.org/project/langdetect/ version 1.0.8

[3] https://spacy.io/models/ro

Further, we removed artefacts resulted from the scraping process: &amp, &gt, &lt, and also deleted the next line carrier symbol.

3.4. Dataset split

We provide train, validation and test splits of the dataset for future model training & evaluation. However, the multi-label nature of the dataset posed problems in balancing the splits so they contain an overall even distribution of labels. As shown in (Szymański and Kajdanowicz, 2017), traditional single-label approaches to stratifying data fail to provide balanced datasets. An imbalanced train/validation/test dataset might include large differences of certain labels(emotions) between splits, leading to poor classifier performance. To overcome this problem we used the implementation of iterative stratification from (Szymański and Kajdanowicz, 2017) which managed to provide a very balanced label count for each emotion, with overall differences of less than 5% per label between splits.

The train split contains 75% of tweets, validation 10% and the test split containing the remaining 15%.

4. Dataset Analysis

4.1. Inter Agreement Score

Classical methods for computing the Inter Agreement Score such as Cohen's kappa (Rosenberg and Binkowski, 2004), or Krippendorff's alpha (krippendorff, 2011), were not suitable either due to the fact that they were not designed for multi-labeled data or because the condition that every annotator should rate the entire dataset was not met in our case.

For this reason, we computed a different score in order to determine the agreement of annotators per tweet, and we also show the results of this score on a simulated dataset with a uniform distribution of labels on 5449 texts.

The initial hypothesis for computing the IAA score is that all annotators have the same level of expertise. We will further call them experts.

We have to classify N texts into K classes, with labels $c_1, c_2, ..., c_K$. Expert 1 thinks that the text x would fit into class $c_{J1}(x)$, where $J_1 \subset 1, 2...K$ is a set of indices. Second expert thinks that the text x would rather fit into class $c_{J2}(x)$, where $J_2 \subset 1, 2...K$, and so on, up to expert m who thinks that the same text x would rather fit into class $c_{Jm}(x)$, where $J_3 \subset 1, 2...K$.

We take into consideration the agreement between experts on both the case when they select the same label per tweet, and the case when they both decide not to select a particular label. The IAA score per tweet will be computed by counting the number of agreed upon labels by annotator J_i and annotator J_j in annotation matrix (both annotators agreed on either selecting or not selecting a particular label), using the following formula:

$$\beta(J_1, ...J_m) = 1 - \frac{2}{K * m(m-1)} \sum_{1 \le i \le j \le m} |J_i \Delta J_j| \quad (1)$$

where m is the number of experts deciding upon the text, K is the number of labels, and $|J_i \Delta J_j|$ is the number of elements in the symmetrical difference between set J_i and set J_j, and is computed with the following formula:

$$|J_i \Delta J_j| = |J_i - J_j \cup J_j - J_i| \quad (2)$$

We use the notation $|A|$ to express the number of elements in set A.

In our annotation setup we have 3 experts and 7 labels (Sadness, Surprise, Fear, Anger, Neutral, Trust, Happiness). Thus, in our case, Equation 1 becomes:

$$\beta(J_1, J_2, J_3) = 1 - \frac{|J_1 \Delta J_2| + |J_1 \Delta J_3| + |J_2 \Delta J_3|}{3K} \quad (3)$$

Based on Equation 2, we compute inter-agreement scores for each text. The resulted mean is 0.84, and the median is 0.8. The histogram is shown in figure 2.

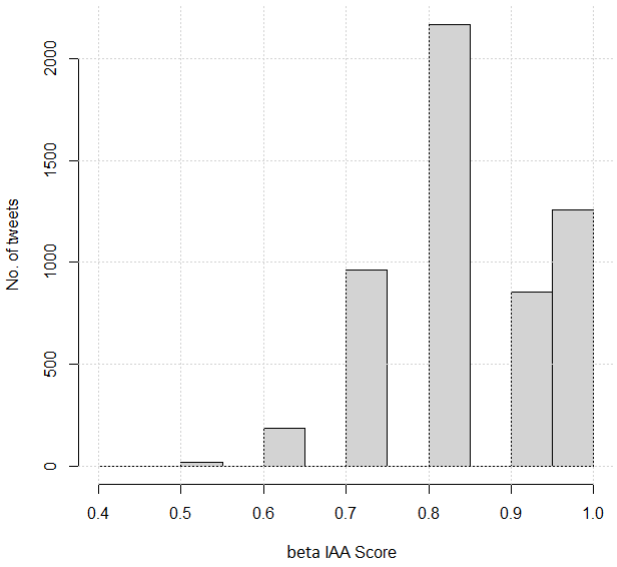

Figure 2: Histogram of REDv2 tweets using β IAA score.

Figure 3 shows the histogram of a simulated dataset with a uniform distribution of labels, where we compute the β score on each tweet.

Also, in our case, with m=3, m being the number of annotators, the minimum value of β is 1/3. In Figure 2 the minimum value of β is 0.4 because of the fact that prior to computing IAA scores on the dataset we eliminated the tweets having not even a partial agreement on the labels by at least two annotators out of three.

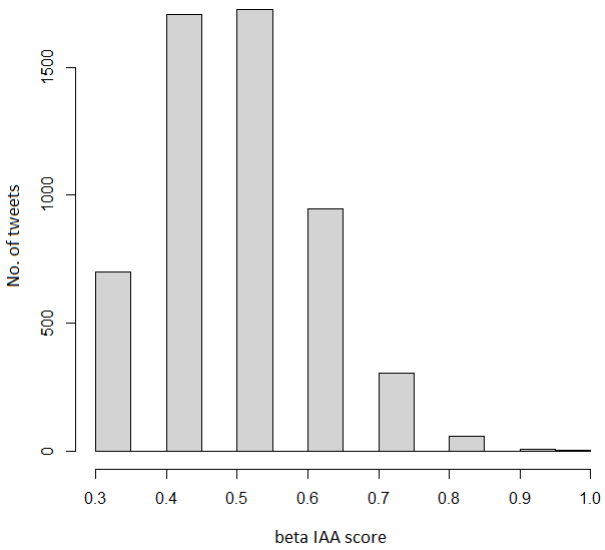

Figure 3: Histogram of random labelled tweets using β IAA score.

In Table 4 we show the statistical summary on the results of β IAA score on both REDv2 and the simulated dataset. Considering the random distribution of labels case where the mean IAA score computed with Equation 3 is 0.5, our result of 0.84 would mean that the dataset is annotated with good confidence. This statement is also reinforced by the fact that we can see in the histogram of the random generated dataset (Figure 3) that there are no texts with perfect agreement of 1, in contrast to the histogram of REDv2 (Figure 2) where we can see that more than 1000 texts have perfect agreement.

	Mean	Median
Random Simulation	0.4951	0.5238
REDv2	0.8440	0.8095

Table 4: Summary of β applied on the simulated dataset with random labels vs on REDv2.

4.2. Setting the Ground Truth

The next step was to set the ground truth of the texts by combining the annotations coming from the three annotators. The dataset comes with two settings of the annotators' agreement per each tweet: the *classification* setting, and the *regression* setting.

4.2.1. The classification setting

This particular way to settle the different annotations for each tweet was created by setting a label only if at least 2 annotators agreed on it. For example, if only one annotator of the three felt that a tweet was sad, the label *Sad* was not set.

The resulting label distribution is shown in table 5, where 87.25% of tweets from the dataset have been labelled with one agreed emotion, 12.31% of tweets with two emotions and the rest labelled with 3 or 4.

No. of labels	No. of tweets	Percentage
1	4754	87.25
2	671	12.31
3	23	0.42
4	1	0.02

Table 5: Percentage of tweets by number of labels for categorical setting

4.2.2. The regression setting

The task of creating an emotion detection model can also be viewed as a regression problem, by taking into consideration all labels received by a tweet, with their corresponding degree of appearance in the annotation matrix.

First, we summed the lines in the annotation matrix so that each tweet received as labels the sum of the labels put by all the three annotators, according to the formula:

$$L_t = (L_{t,A_1} + L_{t,A_2} + L_{t,A_3})/3 \quad (4)$$

where L_t is the set containing the labels for tweet t and L_{t,A_x} is the set of labels given by annotator A_x.

Thus, each tweet receives as final label a vector that can contain one or more of the following values: 0, 0.33, 0.66 or 1, where 0 means that none of the annotators agreed on the label corresponding to the first position in vector, 0.33 means that one annotator out of three selected the label, 0.66 means that two annotators out of three selected the label, and 1 means that all three annotators agreed on the label.

In this setting, if we eliminate duplicate labels per tweet, 31.9% of tweets are single-label, 39.66% of tweets have two labels, 22.92% three labels, and the rest of the tweets, 5.6%, have 4 labels or more. The detailed number of tweets and their corresponding number of labels are shown in Table 6.

No. of labels	No. of tweets	Percentage
2	2908	39.66
1	1339	31.90
3	1681	22.92
4	359	4.9
5	42	0.57
6	3	0.04
7	1	0.01

Table 6: Percentage of tweets by number of labels in the regression setting

In figure 4 we show the correlation of emotions for the regression setting using the corrplot R package[4]. We

[4] https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html

can see that we have the most positive correlations between Happiness and Trust, Sadness and Anger, Fear and Sadness, while the most negative correlations are between Sadness and Happiness, Anger and Happiness, as well as Fear and Happiness. The results are in correspondence with common sense.

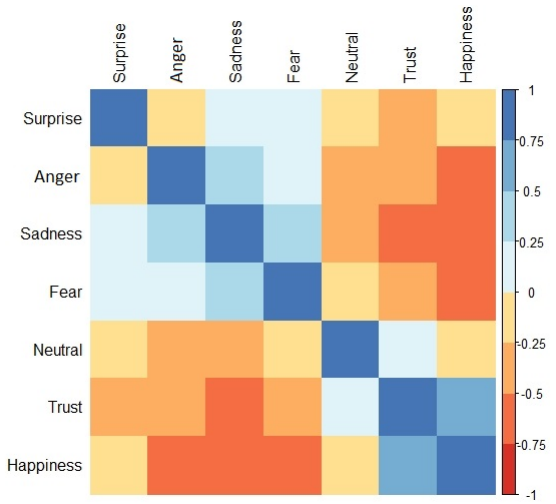

Figure 4: Correlation of emotions by common appearances

5. Experiments and Results

In this section we set a baseline for the REDv2 dataset using transformer models.

As previously specified, given the fact that the dataset has 3 annotators per tweet, we can formulate the problem of emotion detection either as the task of multi-label *classification* per tweet using the agreement labels, or as the task of *regression* on the mean values of the 3 annotators' labels. Whereas in the first setting we will obtain a True/False prediction per emotion given a tweet, in the second setting we will obtain a percentage of how likely each emotion is reflected in the respective tweet.

The dataset is evaluated using two models, the monolingual Romanian BERT (Dumitrescu et al., 2020) and the multilingual XLM-Roberta (Conneau et al., 2019), in both problem settings.

5.1. Model architecture

The model itself is straight-forward:

1. The input of the model is the tokenized tweet, going directly in the transformer which outputs a sequence of token embeddings
2. The embeddings are averaged and a fixed 0.1 dropout is applied to the resulting vector
3. The vector goes into a dense layer that outputs 7 neurons, corresponding to the 7 labels, followed by a sigmoid non-linearity.

	Ham	Acc	F1	MSE
Ro-BERT[1]	0.104	0.541	0.668	26.74
XLM-Roberta[1]	0.121	0.504	0.619	18.41
Ro-BERT[2]	**0.097**	**0.542**	**0.670**	10.06
XLM-Roberta[2]	0.104	0.522	0.649	**9.56**

Table 7: Romanian BERT and XML-Roberta in the classification setting ([1]) and regression setting ([2]). *Ham* represents the Hamming loss

4. Depending on the type of training desired, two loss functions are present: the Binary Cross Entropy loss for the multi-label problem formulation, and the Mean Squared Error - MSE (Wiki, 2022b) loss for the regression problem formulation

We note that every tweet on which the model trains contains both the categorical labels and the mean annotator scores. Thus, when the model outputs the 7 logits passed through the sigmoid, we use this value for the regression setting, and the thresholded value at 0.5 for the categorical setting. Thus, in both settings we can compute the MSE loss as well as the Hamming loss, accuracy and F1 scores.

5.2. Evaluation

Training is performed on the train dataset, with early stopping on the validation set. Results are reported on the test set. Both categorical and regression settings are ran with the same hyperparameters: patience = 5, batch size = 16 and learning rate = $2e^{-5}$. The models smoothly converge within a few epochs.

Before discussing results, for our multi-labeling problem, we choose the Hamming loss (Wiki, 2022a) as the most representative metric, as it is commonly used in such settings and it also accounts for label imbalance. The Hamming loss is the fraction of wrong labels to the total number of labels; it penalizes only the individual labels. We also report the accuracy score which is simply how many times has the model predicted *all* emotions correctly. The F1 score is actually the *micro*-F1, computed by counting the total true positives, false negatives and false positives. The Hamming, accuracy and F1 scores are computed on the predicted binary labels, while the MSE is computed on the logits directly after applying the sigmoid non-linearity to bring them to the [0-1] value interval.

Table 7 shows the summary of the baselines, while table 8 shows a breakdown of the best performing model results per label.

Looking at the results of table 7, our intuition is validated: the Romanian BERT, while less than half the size of the XLM-Roberta (124M parameters vs 278M) performs overall better. This performance also correlates with the Word Fertility Rate computed on the dataset: BERT uses, on average, 1.39 tokens to encode each word in the tweets, while XML-Roberta uses 1.67 tokens per word. More tokens per word leads to longer

	Acc	F1	P	R
Anger	0.88	0.69	0.62	0.77
Fear	0.92	0.63	0.59	0.66
Happiness	0.92	0.65	0.56	0.77
Sadness	0.91	0.75	0.74	0.76
Surprise	0.81	0.59	0.53	0.67
Trust	0.91	0.54	0.45	0.68
Neutral	0.93	0.78	0.73	0.83

Table 8: Breakdown of Romanian BERT per-label metrics in the regression setting

sequences that yield less distinct values after averaging all the tokens' embeddings.

Another interesting result is why did optimizing the MSE objective led to slightly better results than the Binary Cross Entropy which is the standard loss in multi-labeling problems? One answer could be that modeling the mean annotators' values rather than just the most agreed-upon labels actually models the predicted probability distribution of the labels closer to the true distribution; there is less information loss in the mean values than in the thresholded values. However, results are very close in both settings. Finally, while XLM-Roberta did obtain the lowest MSE error, it did not obtain the best Hamming loss nor accuracy.

Looking at the breakdown per label, accuracy is generally high due to the number of zeros in the dataset, as accuracy computes the true positives (*TP*) and true negatives (*TN*) divided by the number of instances. F1 is lower as it is a function of precision (*TP/TP+FP*) and recall (*TP/TP+FN*). We generally see higher recall than precision.

Table 8 also shows that emotions Trust and Happiness are generally easier to predict than the others.

Overall, we can draw a few conclusions: (1) looking at the individual emotions' detection rate, accuracy is generally high, thus validating this dataset as usable in industry; (2) there is sufficient room for improvement as shown by the overall micro-F1, thus this dataset will not be saturated by near-human level accuracy of stronger models in the near-future, lending to much needed research and development in this area for the Romanian language.

6. Conclusions and Future Works

This article presents REDv2, an enhanced emotion detection dataset containing 5449 tweets multi-labeled with 7 emotions: anger, fear, happiness, sadness, trust, surprise and neutral.

The dataset creation procedure has been described and analyzed. Given our annotating constraints, we propose a new IAA score, that, to our knowledge has not been used in the literature so far. We document this score in detail and demonstrate its validity, as well as the dataset's reliability, by comparing statistical summaries and histograms on both our dataset, and on a similar but randomly created dataset, using a normal distribution of labels. The results show an overall reliability of REDv2 of 0.82, versus 0.52 obtained on the random dataset.

We provide train/validation/test splits of the dataset, created with an iterative stratification strategy to provide an overall balanced label distribution between the splits.

Furthermore, building upon the fact that we have 3 annotators per tweet, we provide 2 types of final annotations: a *categorical* one, created by keeping an emotion only if at least 2 of the 3 annotators agreed upon it (thus labeling a tweet with 7 binary values for the 7 emotions), and a *regression* one, created by averaging the 3 annotations for each emotion (labeling the same tweet with a 7-valued array of numbers between 0 and 1).

Finally, we propose strong baselines with two transformer models. We compare a monolingual Romanian BERT model versus a multilingual Roberta model, showing that even if Roberta is twice the size, it still obtains slightly worse results than the Romanian BERT. The two models were tested in both categorical and regression settings, with the regression setting obtaining the best performance, and we discuss around the findings. We also note per-emotion accuracy, showing that some emotions are more easily recognizable than others.

For a proposed REDv3 we plan to further increase the size of the dataset, to add disgust and anticipation classes and to manually check the invalid labeled texts to increase confidence in the annotated tweets, leading to a direct increase of the emotion detection models' performance.

7. Acknoledgements

We would like to thank the students from Cognitive Science, Faculty of Psychology and Educational Sciences, University of Bucharest, for their work done annotating the texts, as well as Nicu Ciobotaru, Gheorghiță Zbăganu and Ana-Sabina Uban for insightful conversations that led to the final form of this article.

8. Bibliographical References

Al-Rubaiee, H., Qiu, R., and Li, D. (2016). The importance of neutral class in sentiment analysis of arabic tweets. *International Journal of Computer Science and Information Technology*, 8:17–31, 04.

Briciu, A. and Lupea, M. (2017). Roemolex - a romanian emotion lexicon. *Studia Universitatis Babeș-Bolyai Informatica*, 62:45–56, 12.

Canales, L. and Martínez-Barco, P. (2014). Emotion detection from text: A survey. In *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*, pages 37–43, Quito, Ecuador, October. Association for Computational Linguistics.

Ciobotaru, A. and Dinu, L. P. (2021). Red: A novel dataset for romanian emotion detection from tweets. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 296–305, Varna, Bulgaria, September. INCOMA Ltd.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). Goemotions: A dataset of fine-grained emotions. pages 4040–4054, 01.

Dinu, L. P., Iordache, I.-B., Uban, A. S., and Zampieri, M. (2021). A computational exploration of pejorative language in social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3493–3498, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Dumitrescu, S. D. and Avram, A.-M. (2019). Introducing ronec–the romanian named entity corpus. *arXiv preprint arXiv:1909.01247*.

Dumitrescu, S. D., Avram, A. M., Morogan, L., and Toma, S.-A. (2018). Rowordnet – a python api for the romanian wordnet. In *2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–6.

Dumitrescu, S. D., Avram, A., and Pyysalo, S. (2020). The birth of romanian BERT. *CoRR*, abs/2009.08712.

Gievska, S., Koroveshovski, K., and Chavdarova, T. (2015). A hybrid approach for emotion detection in support of affective interaction. *IEEE International Conference on Data Mining Workshops, ICDMW*, 2015:352–359, 01.

Huang, C., Trabelsi, A., and Zaïane, O. (2019). Seq2emo for multi-label emotion classification based on latent variable chains transformation, 11.

Iglesias, C. and Moreno, A. (2019). Sentiment analysis for social media. *Applied Sciences*, 9:5037, 11.

Istrati, L. and Ciobotaru, A., (2022). *Automatic Monitoring and Analysis of Brands Using Data Extracted from Twitter in Romanian*, pages 55–75. 01.

Jabreel, M. and Moreno, A. (2019). A deep learning-based approach for multi-label emotion classification in tweets. *Applied Sciences*, 9:1123, 03.

krippendorff, k. (2011). Computing krippendorff's alpha-reliability. 01.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. 01.

Lupea, M. and Briciu, A. (2019). Studying emotions in romanian words using formal concept analysis. *Computer Speech Language*, 57, 03.

Mohammad, S. and Bravo-Marquez, F. (2017). Emotion intensities in tweets. pages 65–77, 01.

Mohammad, S. and Kiritchenko, S. (2018). Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Mohammad, S., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018). SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana, June. Association for Computational Linguistics.

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., and Liang, X. (2018). doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Peng, S., Cao, L., Zhou, Y., Ouyang, Z., Yang, A., Li, X., Jia, W., and Yu, S. (2021). A survey on deep learning for textual emotion analysis in social networks. *Digital Communications and Networks*, 10.

Plutchik, R. (1973). The nature of emotions. pages 810–817, 01.

Rosenberg, A. and Binkowski, E. (2004). Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points. 01.

Scherer, K. and Wallbotth, G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66:310–328.

Strapparava, C. and Mihalcea, R. (2007). SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic, June. Association for Computational Linguistics.

Szymański, P. and Kajdanowicz, T. (2017). A scikit-based Python environment for performing multi-label classification. *ArXiv e-prints*, February.

Szymański, P. and Kajdanowicz, T. (2017). A network perspective on stratification of multi-label data. In Paula Branco Luís Torgo et al., editors, *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, volume 74 of *Proceedings of Machine Learning Research*, pages 22–35. PMLR, 22 Sep.

Tache, A., Gaman, M., and Ionescu, R. T. (2021). Clustering word embeddings with self-organizing maps. application on laroseda – a large romanian sentiment data set. 01.

Wiki. (2022a). Hamming distance, Apr.

Wiki. (2022b). Mean squared error, Apr.

9. Language Resource References

Alexandra Ciobotaru and Liviu P. Dinu. (2021). *Romanian Emotion Detection Dataset*. ISLRN 582-943-832-281-0.