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]
[1]University of Bucharest, Faculty of Mathematics and Computer Science, [2]Independent Researcher, [3]Adobe

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.

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, 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 REDv1 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.

Table 1. Sample annotated texts from REDv2, with English translations.

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

Dataset

Our dataset consists of 5449 tweets, labelled for one or more of the following emotions: sadness, surprise, fear, anger, trust, happiness, or neutral (7 labels).

We scrapped tweets using *query words* from RoEmoLex (Briciu and Lupea, 2017), for the new classes, trust and surprise, in the time-frame 1st of February 2020 - 1st of February 2021.

Table 2. No. of query words per class in REDv1 & REDv2

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 3. Total no. of tweets after 1st annotation step

Class Name	No. of tweets
Anger	1336
Fear	1406
Happiness	1186
Sadness	1299
Surprise	726
Trust	1145
Neutral	852

The **First Annotation Step** 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 single-label dataset, containing **7947** annotated tweets, with the labels: sadness, happiness, fear, anger, surprise, trust and neutral.

The **Second Annotation Step**, which leaded to the final multi-label version of REDv2, involved 66 annotators, also psychology students whose primary language is Romanian, who received sections of the dataset (360-370 tweets) to annotate using Doccano, having the possibility to mark tweets as invalid as well. The sectioning was done so that each of the 7947 unique tweets was assigned to and annotated by 3 annotators.

Dataset Preprocessing and Split. We removed tweets marked as invalid by at least one annotator, and those tweets having full mismatch between annotators, resulting a dataset containing 5449 tweets. We masked proper names using entity recognition (Dumitrescu and Avram, 2019), as well as emails, hyperlinks, usernames, telephone numbers and email addresses using regex methods. The split was done using iterative stratification (Szyma´nski and Kajdanowicz, 2017), of 75% tweets for training, 10% tweets for validation and 15% tweets for testing.

Inter Agreement Score

While classical methods were not suitable for our set-up, we computed our own: 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 $J_m \subset 1, 2...K$.

The IAA score will be computed using the formula: $\beta(J_1, ...J_m) = 1 - \frac{2}{K * m(m-1)} \sum_{1 \le i \le j \le m} |J_i \Delta J_j|$

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: $|J_i \Delta J_j| = |J_i - J_j \cup J_j - J_i|$

In our annotation setup we have 3 experts and 7 labels.

Thus, in our case, the IAA score is: $\beta(J_1, J_2, J_3) = 1 - \frac{|J_1 \Delta J_2| + |J_1 \Delta J_3| + |J_2 \Delta J_3|}{3K}$

We call this score, the **β** score.

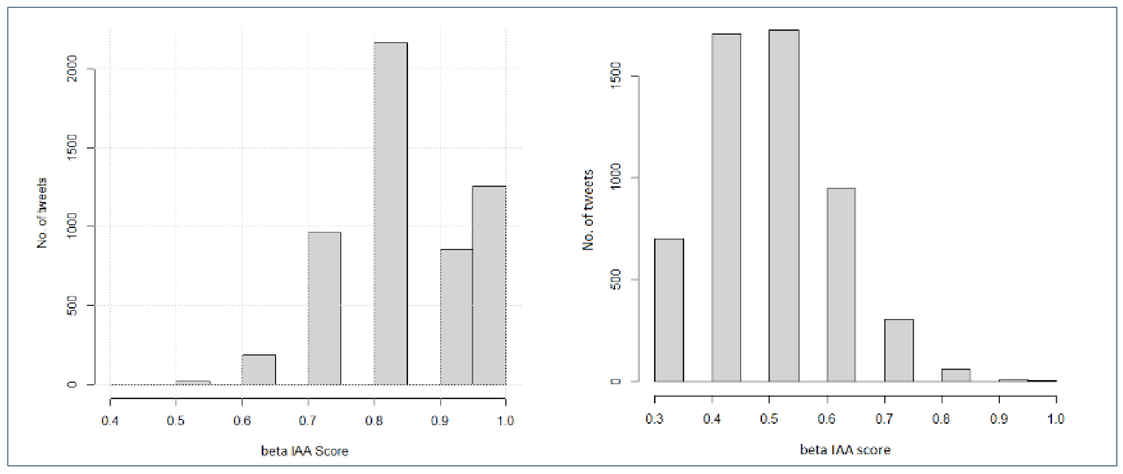

Figure 1. Histogram of REDv2 tweets using **β** score
Mean: 0.84, Median: 0.82

Figure 2. Histogram of random labelled tweets using **β** score
Mean: 0.5, Median: 0.52

Setting the Ground Truth

The Classification Setting. We take into consideration a label only if at least 2 annotators agreed upon it.

The Regression Setting. We take into consideration all labels, with their corresponding degree of appearance in the annotation matrix (Figure 4).

Table 4. Percentage of tweets by no. of labels
Classification Setting

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 no. of labels
Regression Setting

No. of labels	No. of tweets	Percentage
1	2908	39.66
2	1339	31.90
3	1681	22.92
4	359	4.9
5	42	0.57
6	3	0. 4
7	1	0.01

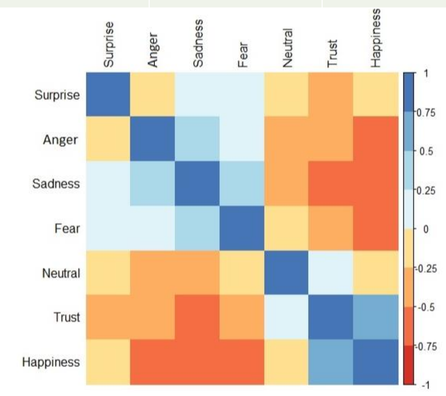

Figure 3. Correlations of emotions by common appearances in the Regression Setting.

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

Ad1 [0 1 0 0 0 0 1]
Ad2 [0 1 0 0 0 0 0]
Ad3 [0 0 0 0 0 0 1]

Figure 4. Annotation matrix with two labels: Surprise and Happiness

Experiments and Results

In the classification setting we obtain 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.

Table 7. Romanian BERT and XML-Roberta in the classification setting ([1]) and regression setting ([2])

Model	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 8. Breakdown of Romanian BERT per-label metrics in the ***Regression Setting***

Class Name	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**

Conclusions and Future Works

- We present REDv2, an enhanced emotion detection dataset containing 5449 tweets multi-labeled with 7 emotions: anger, fear, happiness, sadness, trust, surprise and neutral. We provide two types of annotations, for classification and regression.
- Given our annotating constraints, we propose a new IAA score, that, to our knowledge has not been used in the literature so far, the **β** score, and assess the overall reliability of the dataset (0.84).
- Finally, we propose baselines using two transformer models: monolingual Romanian BERT (Dumitrescu et al., 2020) and multilingual Roberta (Conneau et al., 2019), in both settings.

Selected References

1. 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.
2. Briciu, A. and Lupea, M. (2017). Roemolex - a romanian emotion lexicon. Studia Universitatis Babes, Bolyai Informatica, 62:45–56.
3. Dumitrescu, S. D. and Avram, A.-M. (2019). Introducing ronec – the Romanian named entity corpus.
4. Szyma´nski, P. and Kajdanowicz, T. (2017). A scikit based Python environment for performing multilabel classification.
5. Dumitrescu, S. D., Avram, A., and Pyysalo, S. (2020). The birth of romanian BERT.
6. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzm´an, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale.

© Poster Template by Genigraphics® 1.800.790.4001 www.genigraphics.com