Applying TF-IDF and BERT-based Variants under Multilabel Classification for Emotion Detection in Urdu Language

Sakshi Kalra[1], Saransh Goel[1], Kushank Maheshwari[1], Yashvardhan Sharma[1] and Shresht Bhowmick[2]

[1]*Department of CSIS, BITS Pilani, 333031, Rajasthan, INDIA*
[2]*Greenwood High International School, Gunjur Village, Varthur, Karnataka 560087*

Abstract
Nowadays, the use of emojis is very common to show our emotions with just a single image instead of long sentences describing our emotions. Each emoji describes a particular emotion, such as anger, disgust, fear, sadness, surprise, and happiness. Now if we are given a task to identify emotions in a text, that means we have to tag a text with multiple emojis, each pointing to a different emotion. This paper aims to check for multiple emotions in an Urdu text, which comes under the category of multi-label classification. We have used pre-trained BERT models to add basic knowledge about a language (Urdu in our case). Over the pre-trained model, we added the classification layer using PyTorch. The output layer has seven nodes, six of which are for six emotions, and the seventh is for neutral. FIRE 2022 provided the Urdu tweet dataset used here as part of the subtask ”Multi-label emotion classification in Urdu” of the main task ”Emothreat: Emotion and Threat detection in Urdu.”

Keywords
Social media, UrduHack, BERT, Distil-BERT, Multi-label classification, Negative weight, Positive weight, One vs Rest, Transformers model, Text classification, Tokenizer

1. Introduction

With the vast-scale expansion of social media, it is affecting the narrative of the whole country or even the whole world, which could be evidenced by the examples of various country-wide or worldwide campaigns started from social media accounts and spread into the population. The messages or tweets posted by various users online are responsible for all these new effects of social media, so these messages or tweets must be analysed to understand the mindset of the users about different topics in the public domain. What is better than an emotional classification of text is to categorise it into multiple emotions like anger, fear, sadness, etc. Emotion classification will help in identifying the mood of the population about a topic. This type of task comes under ”affective computing,” as it was defined in [1] in 1995, which is ”computing that relates to, arises from, or influences emotions,” or it can be said that ”affective

FIRE 2022: Forum for Information Retrieval Evaluation, December 9-13, 2022, India
p20180437@pilani.bits-pilani.ac.in (S. Kalra); f20190988@pilani.bits-pilani.ac.in (S. Goel); f20180679@pilani.bits-pilani.ac.in (K. Maheshwari); yash@pilani.bits-pilani.ac.in (Y. Sharma); bhowmickshresht@gmail.com (S. Bhowmick)

computing” is computing that has to do with emotions. [2] Earlier, this was not a famous research field, but nowadays there are hundreds of companies and researchers working on it. Humans express their emotions through a variety of means, including facial expressions, text, audio, body gestures, and movements.Even though our body also physically responds to different emotions by changing heart rate, breathing, etc., for the given dataset, the proposed model is expected to do multi-label emotion detection for text data only. Detecting emotion in text data is not a direct task of identifying some keywords for each type of emotion, sometimes emotion is interpreted through the meaning of the concept and context in a sentence and the interaction between various concepts [3]. The proposed model is based on the same concept of categorising text among different emotions. Text can also contain more than one emotion; for example, someone could be sad and angry at the same time, so this classification becomes a multi-label classification.

The following sections are included in the paper: Section 2 describes the related work; Section 3 describes the dataset and the challenges that go along with it and their solutions; Section 4 describes our model design and techniques; and Section 5 describes the evaluation and result of our model over the data.

2. Related Work

Several authors have participated in the hate speech detection tasks, such as [4], [5], [6], [7], [8] and [9]. Many methods for multi-label classification are used in machine learning, as explained in [10]. The first method is the ranking method. The data is ranked for all classes, and higher ranked classes can be chosen as labels for data points. This is the old method used in machine learning. Other methods include the problem transformation method, in which the multi-label classification is transformed into multiple single-level classifications. This method includes the following steps: (1) randomly select one of the labels for each multi-label instance, (2) discard those instances having multiple labels, (3) consider each different set of labels as a new label, and (4) transform the dataset such that if for a given instance there are three labels, the new dataset will contain three instances of the same data point, making it a multi-class classification problem. One other method is called the algorithm adaptation method; this method includes the custom entropy loss function for multi-label data, the same thing implemented in this paper, and the proposed custom loss function also includes the imbalance part of multi-label classification. [11] describes how we can use the BERT model for text classification. This paper describes the structure of the BERT model, which takes tokenized text as input along with an attention mask and is trained over a large corpus of multilingual data. [2] gives a wholesome survey on emotion detection, finding multi-modal systems to be best for emotion detection tasks. This is the same as humans identifying each other, which is also a multi-modal approach; thus, humans analyse each other’s face, audio, body posture, etc. [3] focuses on emotion detection in text data. The method the paper introduces is called the ”keyword spotting technique” which involves finding some particular keywords as sub-strings from a sentence; each keyword is associated with one or more emotions and can help identify emotion in a sentence. The shortcomings of the keyword spotting technique explained in [3] are that the meaning of a keyword changes with context, such as with the word ”accident,” which is generally associated with a negative

sense, but in this sentence: ”I found my life partner by accident,” the meaning of ”accident” is in a positive sense, so the keyword spotting technique fails in this type of case. [12] experiments with different machine learning based techniques for abusive language detection in Urdu text and achieved an accuracy of 93.6% by using soft voting techniques on three BERT variants (urduhack, BERT and XLM-RoBERTa). The authors in [13] proposed a model for detection of threatening posts using deep learning based models on transformers,they essentially employed the pretrained BERT model (RoBERTa) for classifying text as threatening and non-threatening and obtained an F1 score of 53.46% and ROC AUC of 81.99%.
Another work in[14] fine tuned monolingual and multilingual transformers over Urdu text and used ensembling techniques to combine the results of RoBERTa-urdu-small, XLM-RoBERTa, bert-based-multilingual-case and Alberta-urdu-large and get the accuracy of 0.596 and F1 score of 0.449. The author of [15] got the highest F1 score of 0.7993 by using pre-trained BERT models + fine tuning classification layer over them. They also used data augmentation to make the models generalise better and used both machine learning and deep learning techniques for the task of recognising hate and offensive speech. The effectiveness of several pre-trained multilingual BERT models in the detection of threats and hate speech, which are also types of emotions, is discussed in [14] and [15]. [16] surveys the concept of emotion detection by exploring various methods of categorising emotions one is Direct Emotion Detection, which considers 6 or more basic emotions, believes that all other emotions are a combination of these basic emotions and considers each basic emotion to be independent, whereas Dimensional Emotion Detection, which does not consider emotions to be independent, defines a 2-D or 3-D space for emotion categorisation. The X-axis represents valency, while the Y-axis represents arousal. Each area in the 2-D space shows a certain kind of emotion, and you can also add a Z-axis showing the person’s control over that emotion.

3. Dataset

The dataset is provided by FIRE 2022, under the sub-task A(Multi-label emotion classification in Urdu) of main Task: EmoThreat: Emotion and Threat detection in Urdu. The training dataset contains sentences in Urdu and seven labels for each sentence in one-hot encoding, which is multi-label (0 or 1). the code for this task is available on this[1] github repository. The distribution statistics for each label is as depicted in Table 1 for training data. Each label corresponds towards a particular emotion like anger, disgust, fear, sadness, surprise, happiness, or neutral. The pictorial depiction of training data is as in Figure 1, which on analysis brings about a problem of unbalanced data that could affect the training model parameters in a way that is biased towards labels that have a high number of sentences labelled for it. for example, in the dataset, the label ”neutral data” has a higher count than all others; this problem of unbalanced data is taken care of in further sections by using the method of calculating positive and negative weights for each label.

[1] https://github.com/saransh-goel/emotion$_d$*etection.git*

Table 1
Dataset Statistics

Category	Training Data
Anger	811
Disgust	761
Fear	609
Sadness	2190
Surprise	1550
Happiness	1046
Neutral	3014
Total	9981

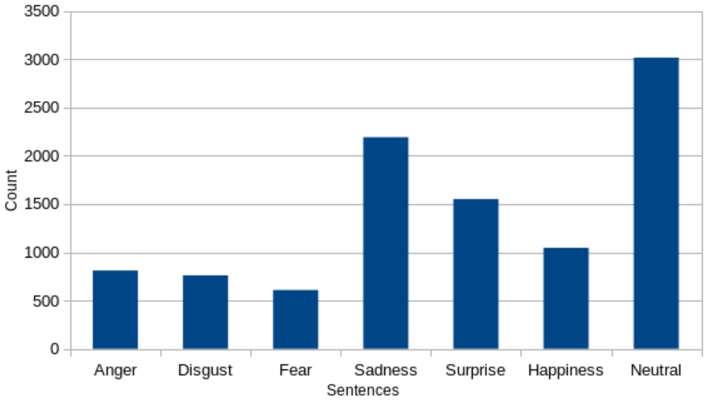

Figure 1: Training set distribution in the Urdu Dataset

4. Proposed Techniques and Algorithms

4.1. Multi-label classification

[17]For a d-dimensional input data, $X \in R^d$ and $Q = \{1, 2, ..., q\}$ set of labels where q is the number of labels. Each instance $x \in X$ can be associated with the subset of labels $L \in 2^Q$ which are called as relevant labels for x and the set of labels in complement of L, i.e $\bar{L} = Q \setminus L$ are called as irrelevant labels for x. Training dataset for multi-label classification of size l, will be a set of elements $(X\, x\, 2^Q)$.

$$i.e., \;\; \{(x_1, L_1), ..., (x_i, L_i), ...(x_l, L_l)\}$$

Multi-label classification would be learning a function $f(x) : X \rightarrow 2^Q$. There are two main methods for multi-label classification:

1. Data decomposition method
2. Algorithm extension method

4.1.1. Data decomposition method

This method includes binary classifiers, One vs Rest method, One vs One method etc. The widely used trick used by this method is to define a function for each class i.e $f_i(x) : X \rightarrow R, i = 1, 2, .., q$ such that $f_k(x) > f_i(x),\ i \neq k\ if\, x \in class\, k.$

$$i.e.,\quad f_k(x) > f_i(x),\ k \in L,\ i \in \bar{L} \tag{1}$$

which means that relevant label should have ranked higher than irrelevant labels.

$$f(x) \;=\; \{k,\ s.t\, f_k(x) \;\geq\; t,\ k = 1, 2, .., q\} \tag{2}$$

Further a threshold can be set for relevant labels, now the methods like one vs rest that the proposed model used came in picture to set this threshold along with the help of binary classifiers like Naive Bayes, SVC and Logistic regression. [17]One vs Rest method divides a q-class multi-label data set into q binary subsets, here the i^{th} subset consists of positive instances with the i^{th} label and negative ones with the all other labels. This method helps in identifying the threshold t in eq 2.

4.1.2. Algorithm extention method

This method includes using multi-class classifiers and dealing with multi-label classification in one function only, like in the proposed model, which uses the BERT pre-trained model with a classification head over it for the multi-label classification.

The BERT variants used as base models are UrduHack and distil-BERT, which can work with multilingual data as the dataset contains sentences in Urdu. For multi-label classification, as shown in figure 2, in the proposed model there are seven parallel feed forward dense layer networks for each class. Each gives a two-node output and works as a binary classifier for its own class.The training process begins with data tokenization and padding with the required [CLS] and [SEP] tokens, followed by passing the tokenized text as input to the Bert model and using the output from the Bert model corresponding to the [CLS] token as input to classification layers as described in [11]. Table 2 lists the various hyperparameters used while training the proposed model.

Table 2
Various Hyperparameters and its Descriptions

Hyperparameter	Description
Learning Rate	1e-05
Number of Epochs	4
Batch Size	2

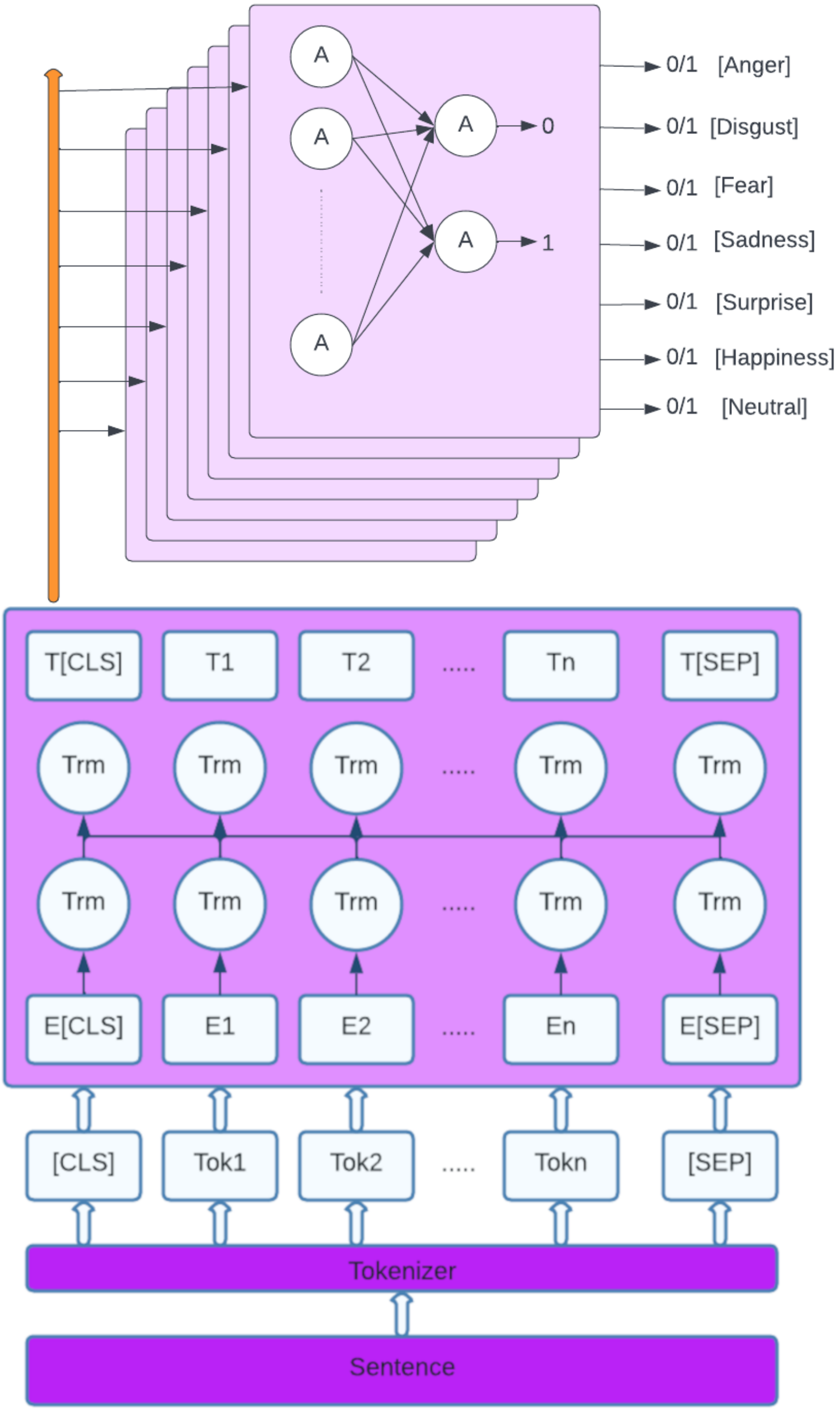

Figure 2: The Proposed Architecture for BERT

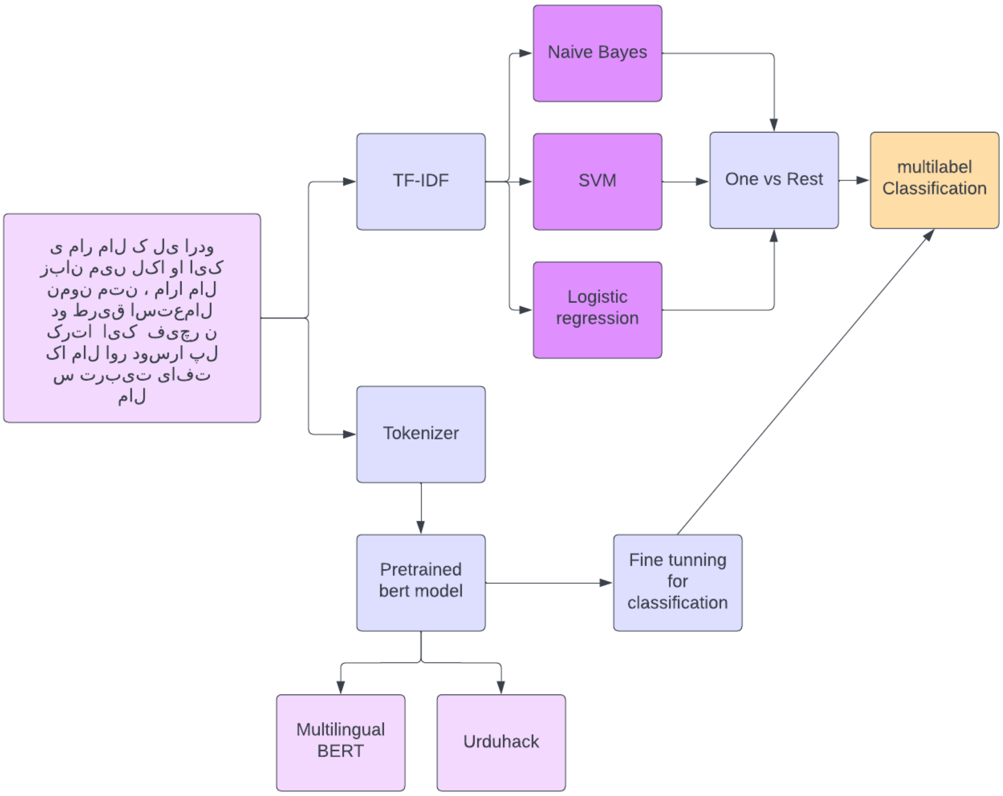

Figure 3: The Proposed Model

4.2. Handling the class imbalance issue

As it is clear from the figure 1 that data is imbalanced, so we firstly calculate negative and positive weights of all classes as follows:

$$Positive\,Weight = \frac{1}{no\,of\,1's\,for\,that\,class} * \frac{total\,number\,of\,data\,points}{total\,number\,of\,classes} \tag{3}$$

$$Negative\,Weight = \frac{1}{no\,of\,0's\,for\,that\,class} * \frac{total\,number\,of\,data\,points}{total\,number\,of\,classes} \tag{4}$$

Then we added these weights in our custom loss function which is calculating cross entropy loss for each class separately and multiplying it with corresponding weights:

$$Loss = (positive\,weight) * (y_{true} * log(y_{pred})) + (negative\,weight) * ((1 - y_{true}) * log(1 - y_{pred})) \tag{5}$$

and added all the losses for each class to form a final loss for a data point.

$$Total\,Loss = \sum_{c=1}^{7} loss(c) \tag{6}$$

5. Evaluation and Results

[10]As this is multi-label classification, so the formula for calculation of accuracy, precision and recall will change and the modified formulas are as follows: If D is the dataset, H is the model, Y are the real labels, Z are the predicted labels(Z=H(D))

$$Accuracy(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \tag{7}$$

$$Precision(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cap Z_i|}{|Z_i|} \tag{8}$$

$$Recall(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cap Z_i|}{|Y_i|} \tag{9}$$

$$F1(H, D) = \frac{2 * precision(H, D) * recall(H, D)}{precision(H, D) + recall(H, D)} \tag{10}$$

The performance of each model is evaluated using various evaluation metrics. Table 3 lists the accuracy, precision, recall, and F1-measure using the TF-IDF model. Table 4 lists BERT variants. In the TF-IDF feature extraction method, linear SVC performed best for emotion detection, followed by logistic regression as the second-best method, and Naive Bayes as the worst. Among BERT variants, UrduHack performs better than multilingual BERT.

Table 3
Performance Evaluation using BERT Variants

BERT Variants	Accuracy	Precision	Recall	F1-Measure
Urdu-hack	0.394	0.453	0.758	0.567
Multilingual-BERT	0.340	0.358	0.8559	0.5048

Table 4
Performance Evaluation using TF-IDF feature extraction

Classifiers	Accuracy	Precision	Recall	F1-Measure
Naive-Bayes	0.68	0.98	0.68	0.80
LinearSVC	0.79	0.91	0.84	0.87
Logistic Regression	0.77	0.96	0.79	0.86

6. Conclusion and Future Work

The proposed results demonstrate that the TF-IDF feature extraction model works better than the BERT model. This is because in emotion detection, keywords are found to be more important than context, as each emotion has its own set of keywords that help a lot with classification. This paper only deals with emotion detection in text data, but as explained in one of the earlier sections, a multi-modal approach to emotion detection is very effective as other features other than text, such as audio pitch and facial expression, more clearly explain an emotion. For image data, an expression detection model could help identify different emotions, and just as each emotion has its own set of expressions, sometimes the same sentence has different meanings with different expressions, such as ”he is very intelligent.” This sentence with a good expression will come in the category of happiness and praise, but with an expression of sarcasm, it will come under the category of jealousy. The same is true for audio, where pitch can help distinguish between anger, excitement, and a lazy tone.

References

[1] J. Oliver, B. García-Zapirain, Affective computing and education, in: INTED2017 Proceedings, IATED, 2017, pp. 1334–1338.

[2] J. M. Garcia-Garcia, V. M. Penichet, M. D. Lozano, Emotion detection: a technology review, in: Proceedings of the XVIII international conference on human computer interaction, 2017, pp. 1–8.

[3] S. N. Shivhare, S. Khethawat, Emotion detection from text, arXiv preprint arXiv:1205.4944 (2012).

[4] S. Butt, M. Amjad, F. Balouchzahi, N. Ashraf, R. Sharma, G. Sidorov, A. Gelbukh, Overview of EmoThreat: Emotions and Threat Detection in Urdu at FIRE 2022, in: CEUR Workshop Proceedings, 2022.

[5] S. Butt, M. Amjad, F. Balouchzahi, N. Ashraf, R. Sharma, G. Sidorov, A. Gelbukh, EmoThreat@FIRE2022: Shared Track on Emotions and Threat Detection in Urdu, in: Forum for Information Retrieval Evaluation, FIRE 2022, Association for Computing Machinery, New York, NY, USA, 2022.

[6] N. Ashraf, L. Khan, S. Butt, H.-T. Chang, G. Sidorov, A. Gelbukh, Multi-label emotion classification of urdu tweets, PeerJ Computer Science 8 (2022) e896.

[7] L. Khan, A. Amjad, N. Ashraf, H.-T. Chang, A. Gelbukh, Urdu sentiment analysis with deep learning methods, IEEE Access 9 (2021) 97803–97812.

[8] I. Ameer, N. Ashraf, G. Sidorov, H. Gómez Adorno, Multi-label emotion classification using content-based features in twitter, Computación y Sistemas 24 (2020) 1159–1164.

[9] L. Khan, A. Amjad, N. Ashraf, H.-T. Chang, Multi-class sentiment analysis of urdu text using multilingual bert, Scientific Reports 12 (2022) 1–17.

[10] G. Tsoumakas, I. Katakis, Multi-label classification: An overview, International Journal of Data Warehousing and Mining (IJDWM) 3 (2007) 1–13.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[12] S. Kalraa, Y. Bansala, Y. Sharmaa, Detection of abusive records by analyzing the tweets in urdu language exploring transformer based models (2021).

[13] S. Kalraa, M. Agrawala, Y. Sharmaa, Detection of threat records by analyzing the tweets in urdu language exploring deep learning transformer-based models (2021).

[14] S. Kalraa, P. Vermaa, Y. Sharmaa, G. S. Chauhanb, Ensembling of various transformer based models for the fake news detection task in the urdu language (2021).

[15] S. Kalraa, K. N. Inania, Y. Sharmaa, G. S. Chauhanb, Applying transfer learning using bert-based models for hate speech detection (2020).

[16] F. A. Acheampong, C. Wenyu, H. Nunoo-Mensah, Text-based emotion detection: Advances, challenges, and opportunities, Engineering Reports 2 (2020) e12189.

[17] J. Xu, An extended one-versus-rest support vector machine for multi-label classification, Neurocomputing 74 (2011) 3114–3124.