# Latent variable-based multiple instance learning towards label-free polarity detection

Peter Lajos Ihasz
Graduate School of Information Science and Engineering
Ritsumeikan University
Kustasu, Shiga, Japan,
gr0238re@ristumei.ac.jp

Mate Kovacs
Graduate School of Information Science and Engineering
Ritsumeikan University
Kustasu, Shiga, Japan
gr0370hh@ed.ritsumei.ac.jp

Victor V. Kryssanov
Graduate School of Information Science and Engineering
Ritsumeikan University
Kustasu, Shiga, Japan
kvvictor@is.ritsumei.ac.jp

## ABSTRACT
Extracting information from the audio content of the users' dialogic utterances would provide an easily-perturbed set of features that could serve as a reliable and inexpensive mean for emotion recognition, suitable to be applied in commercial software development. Owing to the diversity of audio features, however, emotion recognition in spontaneous dialogues is a complex task, typically requiring the pre-training of classifiers on large collections of labeled data. To escape the necessity of hand labeling, a novel multiple instance learning method is proposed. It performs the bag-label-based instance classification through the extraction of latent variables with variational autoencoders. In this semi-supervised method, the bags themselves are gathered from audio features of "weakly" labeled YouTube videos, thus training is fully automated and does not require manual annotation.

## CCS Concepts
• **Information systems → Clustering and classification**
• **Computing methodologies → Information extraction.**

## Keywords
Multiple instance learning; variational autoencoders; unsupervised clustering; big data.

## 1. INTRODUCTION
Following the line of the previous work [1] and [2], extracting information from the audio content of the users' utterances provides a set of features that could serve as a reliable and inexpensive mean for emotion recognition suitable for commercial software development. By extracting these features from the dialogic context, the interpersonal aspect of verbal utterances can also be analyzed. The latter appears to critically influence the generation and control of the interlocutors' interdependent affective states. Nevertheless, supervised machine learning-based classification of the above features assumes classifiers to be pre-trained on labeled data before they are deployed for real-time recognition. Owing to the diversity of the audio features, emotion recognition in spontaneous dialogues is a complex task, demanding a large amount of labeled data to ensure satisfactory recognition accuracy.


*ICISS 2019*, March 16–19, 2019, Tokyo, Japan

ACM ISBN 978-1-4503-6103-3/19/03…$15.00
https://doi.org/10.1145/3322645.3322654

To escape the necessity of labeling large sets of data for the training of real-time affective classifiers, a novel multiple instance learning (MIL) method is proposed. MIL is a supervised machine learning method, where instead of receiving a set of instances which are individually labeled, the learning algorithm is trained on a set of labeled bags, each bag containing several unlabeled instances. Based on the bags, the algorithm tries to deduce a concept that would label individual instances correctly [3]. The proposed approach utilizes variational autoencoders (VAE) [4] to extract latent variables of the bag-populating audio feature vectors, then selects bag-representative instances through clustering the vectors, based on their position in the latent feature space. Building the proposed MIL on latent variables is to counteract the large amount of noise expected in bags of dialogic utterances. In this semi-supervised method, the bags themselves are gathered from weakly labeled YouTube videos, thus training is fully automated, not requiring any manual annotation.

The rest of the paper is organized as follows. Section 2 describes the conceptual and application background behind the proposed method, which is detailed in Section 3. Section 4 gives implementation details for the experiments conducted, while Section 5 elaborates on the results obtained. Finally, conclusions are formulated and future work is outlined in Section 6.

## 2. BACKGROUND
### 2.1 Conventional Multiple Instance Learning
The standard assumption behind multiple instance learning is that each instance $x \in X$ from the instance space X, has a binary latent label $y \in \{0,1\}$. Thus, $\{x, y\}$ is called an "instance-level concept" where an instance is representing an underlying concept $c \in C$ from the concept space C. A 'bag' is a multiset of instance-level concepts, with instances labeled identical to the target class, called 'positive' labels and instances labeled non-identical, called 'negative' labels. A bag is labeled positive if at least one of its instances has a positive label, and negative if all of its instances have negative labels. The standard assumption is that a bag can be represented by a sole concept [5], and it has been implemented through iterated discrimination algorithms [6]. There are complex problems however, where a bag label is determined by the simultaneous presence of several concepts. Weidmann [7] proposes a two-level classification algorithm to learn multiple concepts.

Most of the MIL methods developed have been applied for image / molecule activity/ document recognition, where recognized entities are the bags themselves. Thus those methods, including the ones described above, are concentrating on the prediction of unseen bags, instead of the prediction of unlabeled instances they contain. In the case of emotion recognition however, MIL would be used to train instance-level predictors based on the bag labels.

## 2.2 MIL for affective analysis

Kotzias et al. [8] inferred labels for each instance through propagating information from the bag labels to the instances by inverting the unknown label aggregation function on the training data. In order to achieve this, the authors used a specific similarity measure to compute a label assignment that is compatible with the group structure of the data, and to simultaneously assign the same label to similar instances. Trained and tested on the Amazon, IMDB and Yelp datasets it classified three sentiments with an accuracy of 86%-88%. As an alternative method that learns to predict the polarity of text segments from bag-level labels, Angelidis and Lapata [9] reduced each segment's class probability distribution $pi$ to a single real- valued polarity score. Class probability distributions $pi$, however, are obtained from a supervised feature map classifier pre-trained on sentence-level labels (on a dataset different from the training data of the study). Trained and tested on the IMDB and Yelp datasets it yielded classification results between 91% and 94% for three polarities.

Although learning methods described above are trained only on a set (bag) of instances (instead of each and every instance), selection and hand labeling of the bags is usually still necessary (in [9] sentence-level labels were also needed for pre-training). In the case of dialogues, section-labeling, with the additional task of reason- based sub-sectioning, is difficult to automate, and is a labor and time demanding task, which can easily nullify the benefits MIL would provide in contrast to segment-level labeling. Accordingly, a method capable of finding sub-sections of dialogues, applicable as emotion-related bags needs to be developed. This paper describes an approach of salvaging YouTube videos in a way that not only makes the videos directly applicable for MIL-based emotion recognition, but also completely eliminates the need for their hand labeling.

# 3. BACKGROUND

## 3.1 Movie scenes as labeled bags

As a vast resource of dialogues, dialogic video scenes (of e.g. movies) can be harvested. YouTube for example, contains millions of videos with dialogic scene contents. These videos are also categorized by the titles they are uploaded with, and several of the videos also contain tags, describing their content. Thus a search query on videos, related to a certain emotion (e.g. 'angry scenes') would supposedly return videos containing dialogic scenes where at some point the emotion of e.g. 'anger' is expressed by at least one of the interlocutors. Furthermore, relevant videos to 'angry scenes' will not only contain verbal and/or non-verbal expressions of anger, but will have them as their base concept. The YouTube 8M dataset (Y8M) [10] contains frame- and video- level audio and image features of 6.1 millions of YouTube videos in total, with thousands of dialogic scene contents (at the present). The dataset is free to download under the Creative Commons Attribution 4.0 International license [11]. Since the original audio- visual versions of the feature-sets in the Y8M are also available and trackable online on YouTube, an indirect YouTube search can be conducted in the Y8M for audio features of emotion-expressing videos.

## 3.2 Instance-level polarity detection

Each feature set selected from the Y8M would contain features of a dialogic scene, focusing on the expression of a certain basic emotion (but probably containing several other basic emotions as well). Thus each feature set would serve as a weakly labeled bag for a certain emotion. As this study thrives to find a method easily applicable for commercial use, the technology sensitive visual features (requiring the application of visual sensors) contained in the Y8M are not processed. From the frame-level and video-level audio features, only frame-level features are to be applied since video-level features would not be applicable for instance-level training. The MIL task is to group the frame-level audio features of each bag into processable instances and automatically label them with basic emotion labels. An instance is to be labeled with the emotion label of the bag (labeled positive in standard MIL) if it represents the concept (a basic emotion) identical to the bag; otherwise it is not labeled (labeled negative in standard MIL). Labeling would be achieved through latent variable-based unsupervised clustering, discussed in the next section. Feature instances of several videos can be grouped together along the emotions they are labeled to represent. To advance real-time emotion recognition through the proposed semi-supervised method, polarity classification assumed to be more applicable than the more fine grained but less reliable emotion recognition.

The bags of emotions, mined from weakly labeled YouTube video features would be merged into positive and negative polarity-representing bags. The polarity of a given emotion would be decided based on Russel's circumplex of emotions [12]. Accordingly, an aggregated set of audio feature instances, affiliated with the same basic emotion would constitute a bag of a certain emotion, and an aggregated set of emotion bags with the same polarity would constitute a final bag of positive or negative polarity.

## 3.3 Unsupervised-clustering based classification

This study proposes a new approach towards multiple instance learning in the form of unsupervised clustering of weakly labeled bag instances, mapped into the feature space of their latent variables. The approach is depicted in Figure 1.

Training: As the first step, frame-level audio feature sets of YouTube videos, edited (by the uploaders of the video) to focus on the expression of a certain basic emotion is to be selected from the Y8M (several sets for each emotion). In particular, the titles of each frame-level audio feature set are to be extracted through the ids attached to the sets and matched to the titles of an online YouTube search. Although only features of the videos referenced in the YouTube 8M dataset are to be utilized in the study, through this method, the synonym- and relevance- measures of the YouTube search would be salvaged.

As the next step, the frame-level audio feature sets from Y8M of each selected video will be concatenated into utterance-level features. The concatenation is conducted based on timestamps provided by an online text converter applied on the online version of the videos.

Extraction of latent variables is achieved through a VAE trained on the utterance-level audio features of all selected videos. The VAE extracts latent variables through encoder layers, transforming the input data into abstract variables. Then, through decoder layers, the variables are transformed back into a predicted input form. During training, the abstract variables are constantly updated according to the loss between the original and predicted input. In VAEs, constraints are added that forces the generation of latent vectors to roughly follow a unit Gaussian distribution. The isotropic Gaussian priors allow each latent dimension in the representation to push itself as far as possible from the other factors. [4] Through training the VAE on all of the features (not separately for feature set), the extracted latent feature space would include all concepts possible for the selected bags.

The concatenated utterance-level features can then be mapped to the latent feature space via transforming each audio feature vector into a vector indicating its affiliation towards each extracted latent variable (i.e. the datapoint's position in the latent feature space). Then all transformed vectors are grouped by the emotions the video (they originate from) was selected by. Thus emotion-bags are created, consisting of utterance-level instances of several videos, supposedly focusing on the expression of the same emotion.

As the final step of the training procedure, the instances are clustered within each bag with an unsupervised clustering method. Once stable clusters are found, instances in the largest cluster are selected as representatives for the given emotion bag. After removing non-representative instances from each emotion bag, the polarity bags can be populated with representative instances of the corresponding emotions. These bags are to serve as the base of a similarity-measure based classification. Emotion bags can also be used before aggregation to train basic emotion classifiers.

Classification: To compile a test set, similarly to the training procedure, first, basic emotion-oriented scenes need to be selected and matched with their frame-level audio feature sets in the Y8M. After concatenating the frame-level features into utterance-level instances, each instance needs to be hand-labeled with basic emotion labels and polarity labels (in accordance with the corresponding emotion label). Next, the VAE pre-trained on the audio features of the training set extracts the latent variables from the audio features of the labeled test instances. Then, the instances can be transformed into vectors representing their position in the latent feature space. Finally, the similarity between the transformed test instances and the instances populating both polarity bags is to be measured to predict the label of each test instance. Comparing the predicted labels with the original hand-made labels yields the classification result.

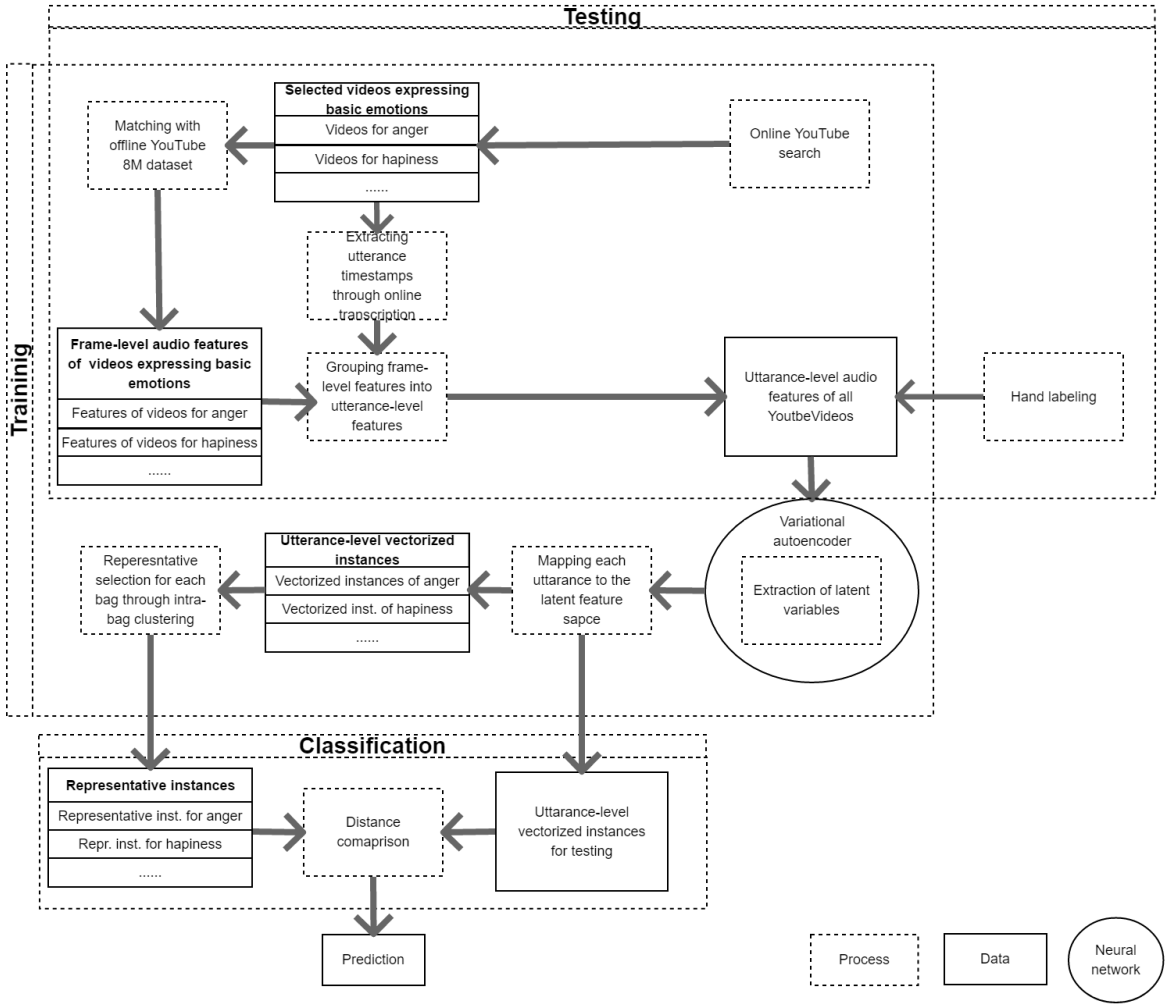


**Figure 1. Latent variable-based unsupervised clustering of audio features extracted from YouTube**

# 4. EXPERIMENTS

## 4.1 Data Mining

As detailed above, frame-level audio feature sets from the YouTube 8M dataset were selected through indirect YouTube search. The search was conducted through the YouTube API [13]. Search- phrases input to the API were subjectively selected English synonyms (based on the Collins Thesaurus [14]) of the eight basic emotions defined by Plutchik [15], which would likely to occur with the word 'scene'. The selected words are: 'happy' for `joy`, 'sad' for 'sorrow', `angry` for 'anger', 'scared' for 'fear', 'surprised' for 'surprise', 'eager' for 'anticipation', 'disgusted' for 'disgust', and 'impressed' for 'acceptance'. Since the word phrases

were in English, the targeted videos were also of English language. Non- English videos with English titles were filtered out from the search results. The selected videos (having an extracted audio feature set in the Y8M) were filtered through the API to be between one and five minutes in length, in order to get videos of dialogues focused on the expression of one particular emotion. For each emotion, 225 minutes of dialogic videos were selected, providing an average of 45 videos per emotions, and an average of 956 utterance-level feature sets retrieved from the Y8M. The utterance-level features were concatenated from frame-level audio features. Each frame contains a 128 dimensional feature vector, extracted from a deep convolutional neural network, trained on log-mel spectrogram patches as described in [16]. 361 videos were selected in total, leading to the extraction of a total of 7650 utterances from the Y8M.

Grouping frame-level instances into utterance-level instances was conducted along time-stamps. Time stamps were provided by the online text converter of Cloud Converter [17], applied on online YouTube video streams of the corresponding videos of the audio feature sets of Y8M.

## 4.2 Data structuring and annotation

Although the training of the proposed method does not require annotation, testing the method for the purpose of the study necessitated the compilation of a test set. 80% of the utterance – level features were used for training the proposed system, while 20% for testing it. In particular, for the 20%, an average of 191 test instances were created for each emotion bag, adding up to 1530 instances in total. The test instances have been annotated by three native English male speakers of age between 27 and 32. The annotators were asked to determine the underlying emotion of the interlocutors for each utterance, while watching the selected YouTube videos online. All utterance-level test instances received one emotion tag. The inter-annotator agreement for emotion tags assessed with Fleiss' Kappa was 69.2%. The emotion labels were then transformed into negative or positive polarity labels based on their valences defined in Russels's circumplex of emotions. The reason for not having the utterances annotated with polarity tags from the beginning is that recognition of specific emotions is also analyzed in the study.

## 4.3 Implementation

### *4.3.1 Latent variable extraction*

The audio data arrays were fed into the encoder layers of the VAE. Since the encoder consists of three consecutive GRU neural network layers ending in a fully connected layer, the decoder also consisted of three GRU layers, where the first layer is input with the output of the decoder. The latent variables are the output of the encoder's fully connected layer. Eight latent variables were extracted under the assumption that in an ideal case each latent variable represents a different emotion (or emotion-related concept that can occur in any bag).

### *4.3.2 Unsupervised Clustering*

Once the eight latent variables got extracted, each emotion bag instance receives the corresponding vector representing its position in the latent feature space. The vectors of each instance were clustered by Expectation Maximization (EM) [18] through Gaussian Mixture Models (GMMs) [19]. GMMs assume that the data points are Gaussian distributed; this is a less restrictive assumption than assuming that they are circular by using the mean (like other clustering methods, such as k-means). Each instance was regarded as being generated by a mixture of Gaussians. To find the parameters of the Gaussians that best explain the data, a conventional EM was used, computing a matrix where the rows are the data point and the columns are the Gaussians [18]:

$$W_j^{(i)} = \frac{\varphi_j N\left(x^{(i)}\right); \mu_j \Sigma_j}{\sum_{q=1}^{k} \varphi_q N(x^{(i)}; \mu_q \Sigma_q)}$$

where $\varphi_j$ is the weight for each Guassian, $\mu_j$ is the mean of the Guassians, and $\Sigma_j$ is the co-variance of each Gaussian. In matrix $W$ an element at row $i$, column $j$ is the probability that $x^{(i)}$was generated by Gaussian $j$. The probability for a given Gaussian is computed in the numerator and normalized along $k$ Guassians in the denominator.

The number of stable clusters were decided based on the Dunn Index cluster validation metric [20] applied through several iterations with varying cluster sizes. The Dunn Index is the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance. It is computed as

$$D = \frac{\min\limits_{c_i \neq c_j \in C} \{d(c_i c_j)\}}{\max\limits_{c_k \in C} \{d'(c_k)\}}$$

where $c_k$ is the number of all clusters in the cluster space $C$ and $d$ is a distance metric chosen to be the Euclidean distance in this study. The Dunn Index has a value between zero and $\infty$, and should be maximized. Once stable clusters are found, the cluster containing the largest number of vectors is selected as a representative cluster for the given emotion. Then, polarity bags are populated with the vectors of the representative clusters of each corresponding (negative or positive valence) emotion.

### *4.3.3 Experimental setups*

To test the efficiency of the latent variable extraction-based classification, three experimental setups were designed with identical architectures but different test and training sizes. The 8:2 train:test ratio was set in all cases. While in Setup 1, the sets used were in the original size of 6120:1530; in Setup 2, the sets were cut to 80% (4896:1224); in Setup 3, the sets were cut to 60% (3672:918). Through the three setups the authors attempted to explore if there is a decreasing tendency with the shrinking of the training sets. The latter, if detected, would prove that the more data provided, the more the autoencoder can learn, which would indicate that: A) to a certain extent, the VAE is learning the right parameters for emotion classification, and that the extracted latent variables are indeed reflecting emotion related abstract concepts. B) a dataset gathered from YouTube videos, even if noisy as it is (in terms of the simultaneous presence of several emotions), is good enough for the purposes of the study.

# 5. RESULTS AND DISCUSSION

Table 1 summarizes the polarity classification results for each experimental setups separately. In the light of the related studies [8], [9], the classification results may appear moderate. Nevertheless, the proposed classifier was trained on only 6120 datapoints maximum in contrast to the above studies, which were trained on hundreds of thousands of instances. The scope of the study allowed only for 1530 hand-labeled test instances, which – to ensure statistical significance of the testing - restricted the authors to use a moderate-size training set. According to the proposed method, however, the training set could be enlarged manifold, limited only by the number of the datasource in use (and the number of instances that can be labeled for testing). As classification accuracy is noticeably higher in experimental setup

2 than in setup 3, and in setup 1 than in setup 2 and 3, it is reasonable to assume that the proposed method would show better performance in accordance with the expected enlargement of the training set.

From the precision, recall and F1 score values, it can be seen that negative polarity has been more accurately predicted with all three setups. The lower recognition accuracy on the positive polarity stipulates that the audio features contain cues that show stronger association with negative emotions.

**Table 1. Polarity classification results**

| | Precision | | Recall | | F1-score | | Overall Accuracy |
|---|---|---|---|---|---|---|---|
| Setup | pos | neg | pos | neg | pos | neg | |
| 1. | 0.52 | 0.85 | 0.63 | 0.76 | 0.58 | 0.81 | 71.22% |
| 2. | 0.59 | 0.75 | 0.63 | 0.71 | 0.62 | 0.74 | 68.03% |
| 3. | 0.40 | 0.61 | 0.48 | 0.53 | 0.44 | 0.57 | 51.13% |

Table 2 further elaborates on the performance of the proposed method through emotion-level classification. The low overall accuracy is a strong argument for choosing the less fine-grained polarity classification to be applied in commercial products. Among all three experimental setups, 'angry' (0.38, in the best performing setup, bps) and 'happy'(0.44, bps) was recognized with the highest F1 score while 'surprised' (0.08, bps) and 'eager' (0.12 bps) with the lowest scores. Since the proposed method shows an especially low recognition accuracy in all three setups with 'eager' and 'surprised', reformulation of search phrases for the basic emotions they refer to ('anticipation' and 'surprise') assumed to be necessary.

**Table 2. Emotion-level classification results**

| | F1-score | | |
|---|---|---|---|
| Emotions | Setup 1. | Setup 2. | Setup 3. |
| Angry | 0.38 | 0.32 | 013 |
| Scared | 0.30 | 0.20 | 0.10 |
| Happy | 0.44 | 0.34 | 0.10 |
| Sad | 0.28 | 0.28 | 0.10 |
| Eager | 0.12 | 0.22 | 0.06 |
| Surprised | 0.08 | 0.10 | 0.05 |
| Impressed | 0.22 | 0.16 | 0.15 |
| Disgusted | 0.36 | 0.30 | 0.12 |
| **Overall acc.** | **29.40%** | **25.58** | **10.51%** |

# 6. CONCLUSIONS

This study proposed a new semi-supervised approach towards polarity detection in dialogues, trained on audio features of emotion-oriented dialogic-scenes from the YouTube 8M database. The training set was not labeled by hand, only "weak" labels of the YouTube search phrases were used through a multiple instance learning approach to automatically annotate the instances. Although the proposed approach needs further investigation and improvements, it yielded promising results even on a small dataset. The approach has the ability to realize training on large sets of unlabeled data, restricted only by the size of the test set (if necessary) and the source of the data. Investigation of other search phrases, as well as other clustering methods and distance measures is left for future work.

# 7. REFERENCES


[1] Vogt, T., André, E., and Bee, N. 2008. EmoVoice. In *Proc. of Int. Tutorial and Res. Workshop on Perception and Interactive Tech. for Speech-Based Sys*. Springer, 188-199.

[2] Fayek H. M., Lech, M., and Cavedon, L. 2015. Towards real-time speech emotion recognition using deep neural networks. In *Proc. of 9th Int. Conference on Signal Processing and Communication Systems (ICSPCS)*. IEEE, 1-5.

[3] Scott, S., Zhang J., and Brown, J. 2005. On generalized multiple-instance learning. *International Journal of Computational Intelligence and Applications* 5, 01, 21-35.

[4] Doersch, C. 2016. Tutorial on variational autoencoders. arXiv:1606.05908. Retr: https://arxiv.org/abs/1606.05908

[5] Babenko, B. 2008. Multiple instance learning: algorithms and applications. PubMed/NCBI Google Scholar. 1-19.

[6] Dietterich, T.G., Lathrop, R. H., and Lozano-Pérez, T. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence* 89, 1-2, 31-71.

[7] Weidmann, N. 2003. *Two-level classification for generalized multi-instance data*. Master's Thesis. Albert-Ludwigs-Universität, Freiburg, Germany.

[8] Kotzias, D., Denil, M., De Freitas, N., and Smyth, P. 2015. From group to individual labels using deep features. In *Proc. of 21th Int. Conf. on Know. Disc. and Data M.*. ACM, 597.

[9] Angelidis, S. and Lapata, M. 2018. Multiple Instance Learning Networks for Fine-Grained Sentiment Analysis. *Trans. of the Asso. of Computational Linguistics.* 6, 17-31.

[10] Abu-El-Haija, S., Kothari, N., Lee, et al. 2016. Youtube-8m: A large-scale video classification benchmark. arXiv:1609.08675. Retr.: https://arxiv.org/abs/1609.08675.

[11] Attribution, Creative Commons. "4.0 International License" http://creativecommons. org/licenses/by/4.0 Retrieved:2018/10/31

[12] Russel, J. A. 1980. A circumplex model of affect. *Journal of personality and social psychology*. 39, 6, 1161-1178.

[13] YouTube API. https://developers.google.com/youtube/v3/ Last accessed:2018/10/31

[14] Collins thesaurus of the English language. https://www.collinsdictionary.com/dictionary/english-thesaurus . Retrieved: 2018/10/31

[15] Plutchik, R. 2001. The Nature of Emotions. *American scientist*. 89, 4, 344-350. DOI: 10.1511/2001.4.344.

[16] Hershey, S., Chaudhuri, S., Ellis, D. P. et al. 2017. CNN architectures for large-scale audio classification. In *Proc. of Int. Conf. on Acoustics, Speech and Signal Proc.* 131-135.

[17] FoxAVideo. Cloud Converter. https://www.360converter.com/ Retrieved: 2018/10/31

[18] Moon, T. K. 1996. The expectation-maximization algorithm. *IEEE Signal processing magazine.* 13, 6,47-60, DOI: 10.1109/79.543975

[19] Reynolds, D. 2015. Gaussian mixture models, Vol 2.: Encyclopedia of biometrics. Springer, 827-832.

[20] Pakhira, M. K., Bandyopadhyay, S., and Maulik, S. 2004. Validity index for crisp and fuzzy clusters. *Pattern recognition*. 37, 3, 487-50