MULTIMODAL BAG-OF-WORDS FOR CROSS DOMAINS SENTIMENT ANALYSIS

Nicholas Cummins[1], *Shahin Amiriparian*[1,2], *Sandra Ottl*[1,3], *Maurice Gerczuk*[3]
Maximilian Schmitt[1,4], *Björn Schuller*[1,4]

[1]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2] Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
[3]Chair of Complex and Intelligent Systems, University of Passau, Germany
[4]Group on Language, Audio, and Music, Imperial College London, UK
nicholas.cummins@ieee.org

ABSTRACT

The advantages of using cross domain data when performing text-based sentiment analysis have been established; however, similar findings have yet to be observed when performing multimodal sentiment analysis. A potential reason for this is that systems based on feature extracted from speech and facial features are susceptible to confounding effecting caused by different recording conditions associated with data collected in different locations. In this regard, we herein explore different Bag-of-Words paradigms to aid sentiment detection by providing training material from an additional dataset. Key results presented indicate that using a Bag-of-Words extraction paradigm that takes into account information from both the test domain and the out of domain datasets yields gains in system performance.

Index Terms— Sentiment Analysis, Cross Domain, Bag-of-Words, Multimodal, Deep Spectrum Features

1. INTRODUCTION

Automatic sentiment analysis is a highly active area of research in affective computing [1, 2, 3]. Moreover, this research has many real world applications [4, 5], e. g. , social media is establishing itself as an alternative channel of political sentiment to conventional polling [3]. For instance there were larger amounts of positive tweets in relation to the Trump campaign compared to the Clinton campaign in the lead up to the 2016 US presidential election [6].

Sentiment is defined as a long term disposition to react emotionally and cognitively towards an object such as a person, product, place or event [3]. Whilst affects are generally expressed through a wide range of verbal and non-verbal communication channels, sentiment analysis has been traditionally focussed on text analysis [1, 2, 3]. However, considerable research efforts have focussed on multimodal sentiment analysis using text, speech and visual cues [7, 8, 9]; for a recent review, the reader is referred to [3]. Of particular interest within multimodal sentiment detection is analysing social media content such as vlogs and reviews posted on sites such as YouTube [9, 10, 11, 12].

A range of different feature representations have been tested for multimodal sentiment detection; for text detection, a variety of approaches have been explored including lexicon polarity, text2vec and bag-of-words representations (See [3] and references within). Standard acoustic feature sets extracted using the openSMILE toolkit are a widely used approach for speech-based sentiment recognition [8, 12]. However, recent results in both sentiment and emotion recognition show the promise of using alternative approaches such as *Deep Spectrum* features for these tasks [13, 14]. Features related to facial expression are widely used for video-based analysis (See [3] and references within), and are well known for capturing non-verbal emotional information [15]. A general trend observed in multimodal sentiment analysis is that text is arguably the most reliably modality; however, improvements in system performance can be gained by adding the other two modalities [3].

This work explores the advantages of adding out-of-domain data when performing multimodal sentiment polarity detection on data collected from YouTube. To the best of the authors' knowledge, while this problem has been explored for text-based sentiment analysis [12, 16, 17], this is the first time such work has been performed for multimodal sentiment analysis. Due to the 'in-the-wild' nature of our data, we use a bag-of-words paradigm to quantise our chosen feature spaces before system training and testing [18]. Bag-of-Words feature representations have recently been shown to outperform state-of-the-art systems such as end-to-end learning in affect detection and similar tasks [19, 20].

The rest of this paper is laid out as follows: Section 2 outlines the datasets used in our analysis and the different feature representations used are explained in Section 3. The key experimental settings are given in Section 4 with the subsequent results given in Section 5. Finally, our concluding remarks and future work plans are given in Section 6.

Table 1. Distribution, mean length and standard deviation of all clips in the Movie Review DataSet between train, devel(opment) and test sets.

Characteristic	Train	Devel	Test	Total
Number of Clips	215	72	72	359
Number of Positive Clips	125	42	42	209
Number of Negative Clips	90	30	30	150
Mean Clip Length (m:s)	2:32	2:27	2:30	2:31
Standard Deviation (m:s)	0:40	0:41	0:44	0:41

2. SENTIMENT ANALYSIS CORPORA

The results presented in this paper are gained on the *Movie Review Dataset* which consists of 359 YouTube videos of individuals reviewing a movie that they have recently watched [12]. Sentiment scores ranging from 1 (very negative) to 6 (very positive) have been assigned to each clip by two annotators. Based on the average sentiment score, the clips are divided into positive and negative classes. Clips having an average sentiment greater than *3.5* are denoted as being *positive* and all other clips are denoted as being *negative*. The corpus is then divided into reviewer independent *Train* (60%), *Development* (20%) and *Test* partitions (20%) (cf. Table 1). For further details on this dataset the reader is referred to [12].

To test if we can achieve higher classification accuracies on the Movie Review Dataset, we supplement the training and development sets with clips from the *Music Review Dataset*. This data set was collected specifically for these experiments using our purpose-built *Cost-efficient Audiovisual Aquisition via social-media small-world Targeting* (CAS^{2}T) toolkit [21]. This dataset contains 237 clips collected from the YouTube channel *theneedledrop*[1]. The clips contain a single individual reviewing music albums giving them a rating between 1 to 10. We convert the ratings into sentiments in the following manner: all clips with ratings less than *6* are assumed to be portraying negative sentiment and all clips with a rating *6* or greater are assumed to be portraying positive sentiment. In our experiments we supplement the Movie Review training set with 177 videos and the development set with a total of 60 videos (cf Table 2). Note that as this dataset contains only one speaker, we do not use it for system testing.

3. BAG-OF-WORDS FEATURE REPRESENTATIONS

To create a common feature space for fusion, the features extracted from the audio, video, and text are quantised into a bag-of-words representation using our openXBOW toolkit [18]. The bag-of-words paradigm quantises 'low-level' features, such as *Mel Frequency Cepstral Coefficients* (MFCC) for audio, local image descriptors for videos or n-grams for text, by assigning each low-level feature to a 'word' from a codebook (dictionary) learnt from some training data. Counting the frequency of assignments for each word generates a fixed length histogram (bag) representation of the underlying feature space.

Table 2. Distribution, mean length and standard deviation of Music Review videos across the train and devel(opment) sets.

Characteristic	Train	Devel	Total
Number of videos	177	60	297
Number of Positive Clips	105	36	177
Number of Negative Clips	72	24	120
Mean Clip Length (m:s)	9:21	9:14	9:19
Standard Deviation (m:s)	3:44	3:09	3:34

It is worth noting here that as our corpora have been collected from YouTube they can be regarded as 'in-the-wild', so are affected by less than ideal recording conditions and other related phenomena. Bag-of-Words representations are well suited for such data; due in part to the quantisation step, they are considered more robust than raw features [19]. This property is particularly important for the video and audio modalities which are susceptible to the influence of noise. For further details on the bag-of-words representations, the interested reader is referred to [18].

3.1. Text Features

Of the three modalities analysed in this work, linguistic feature are arguably the most widely used for sentiment analysis [3, 22, 23]. The Movie Review Dataset is suitably transcribed to allow for straightforward extraction of linguistic features; however, this is not the case for the Music Review Data. We therefore used the *Google2SRT*[2] toolkit to automatically generate transcription of this data. The transcriptions were then stripped of all dispensable content like punctuation and non essential special characters.

When creating the bag-of-words models, we used unigram or back-off bigram sequences, herein denoted as *bag-of-ngrams* (BoNG). Initial tests indicated that suitable BoNG models could be obtained using a minimum term frequency of 10 and a maximum term frequency of 10,000 and applying *inverse document frequency* (idf) weighting.

3.2. Speech Features

We use the recently proposed *Deep Spectrum* features as our low-level audio features. Deep spectrum features have been used for speech-based emotion recognition [13] and sentiment detection [14] and have be shown to produce compa-

[1] https://www.youtube.com/user/theneedledrop

[2] http://google2srt.sourceforge.net/en/

rable performance with more standard speech feature spaces. Deep spectrum features are extracted by feeding audio spectrograms through a pre-trained image classification convolutional neural network (CNN), and using the activations of a fully connected layer as the feature vector [13, 24].

Specifically, we extract power spectrograms over window sizes of 5 s with a hop size of 4 s. The spectrograms themselves are created with Hanning windows of width 16 ms and overlap 8 ms and are plotted and scaled using Python's matplotlibpackage. The spectrograms are then fed through AlexNet [25] obtained from the Caffe [26] model-zoo[3]. The activations of the second to last fully connected layer yield the feature vectors of size 4 096.

When creating the *Bag-of-Audio-Words* (BoAW) representations, we optimise the *codebook size* (Cs) for $Cs \in \{500, 1000, 2000, 5000\}$, and the *number of assignments* (Na) for $Na \in \{1, 10, 20, 50\}$. The input features are normalised to $[0, 1]$ in an online manner (parameters are calculated for the training data and are applied to the test data). Further, the codebook is generated by random sampling [18].

3.3. Visual Features

Facial expressions play a critical role in understanding emotions and sentiments [3, 15]. Using the *OpenFace* toolkit [27], we extracted the following low-level descriptors: gaze, head pose pose, facial landmarks locations (2D and 3D), facial action points. The default OpenFace settings were used and the dimensionality of the resulting feature vector is 427.

When creating the *Bag-of-Visual-Words* (BoVW) representations, we optimised $Cs \in \{2000, 4000, 8000\}$ and $Na \in \{10, 20, 25\}$. As for the BoAW representation, the input features are normalised to $[0, 1]$ in an online manner and the codebook is generated by random sampling of the training data set.

4. EXPERIMENTAL SETTINGS

All results are reported in terms of *unweighted average recall* (UAR) on the development and test set of the Movie Review Dataset. All system parameters are optimised for the development set with the classifiers trained using the training partition(s) only. The classifiers used to generate the test set results are trained using the combined training and development partition(s). We test all modalities individually and in early and late fusion combinations.

Three different experimental set-ups are used to determine the effect of combining data of both Movie and Music Review Datasets when training the different classifiers:

I MOVIE REVIEW ONLY, uses only the Movie Review Dataset in all aspects of the experimental paradigm.

[3] https://github.com/BVLC/caffe/wiki/Model-Zoo

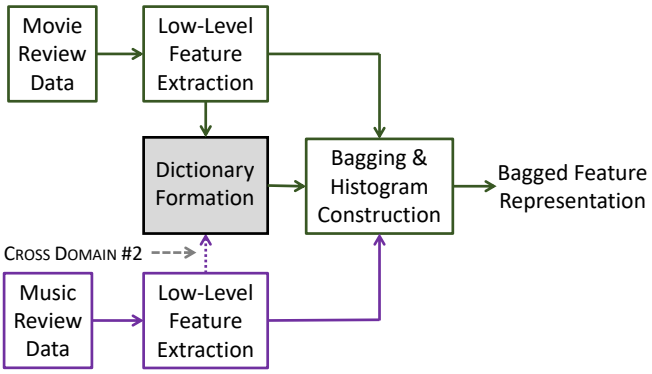

Fig. 1. Two separate method are used to form the cross-corpus bag-of-words features: (i) CROSS DOMAIN #1 in which low-level features from both datasets are bagged with respect to a dictionary created from the Movie Review dataset only; and, (ii) CROSS DOMAIN #2 in which the low-level features from both datasets are bagged with respect to a dictionary created from features from both data sets.

II CROSS DOMAIN #1, both datasets are quantised by codebooks formed from low-level feature vectors taken from the Movie Review Dataset and the resulting bagged features (from both datasets) are used to train the subsequent classifiers (cf. Figure 1).

III CROSS DOMAIN #2: both datasets are quantised by joint codebooks formed from low-level feature vectors taken from both datasets, as in CROSS DOMAIN #1 the resulting bagged features (from both datasets) are used to train the subsequent classifiers (cf. Figure 1).

Classification is performed using a linear SVM with input normalisation applied. The slight imbalance between the two classes (positive and negative sentiment) is counteracted by adjusting the weights for the two classes accordingly, i. e. , 1.4 for negative and 1.0 for positive instances. Training and evaluating is achieved using *WEKA* [28]. The WEKA wrapper of *LibLINEAR* with the *L2-regularized L2-loss* solver is used as the SVM implementation [29].

5. RESULTS AND DISCUSSION

A comprehensive set of tests was done to optimise the bag-of-word set-ups for each modality. For the BoNG, the best development set, 80.7 %, and test set, 79.8 %, were found using a unigrams in the MOVIE REVIEW ONLY set-up (cf. Table 3). In the two cross domain set-ups we observed that back-off bigrams gave the strongest results. However, the addition of the cross domain data did not improve the system performance above the strongest MOVIE REVIEW ONLY scores.

It is worth noting that the performance of bigrams in the MOVIE REVIEW ONLY set-up was 72.4 % for the development set and 73.6 % for the test set. Therefore, it can be ar-

Table 3. Comparison of strongest systems with and without addition of cross domain training data when performing polarity detection sentiment classification. Feature representations are bag-of-ngrams (BoNG), bag-of-audio-words (BoAW) formed from Deep Spectrum features, and bag-of-visual-words (BoVW) formed from facial expression features. The features are also combined by early and late fusion. The performance metric is the unweighted average recall (UAR) and chance level is 50 %.

UAR		MOVIE REVIEW ONLY			CROSS DOMAIN #1			CROSS DOMAIN #2		
		C	dev	test	C	dev	test	C	dev	test
Individual	BoNG	10^{-2}	**80.7**	**79.8**	10^{-2}	76.0	**76.7**	10^{-2}	**78.9**	**79.1**
	BoAW	10^{-4}	69.5	74.5	10^{-7}	71.0	73.3	10^{-6}	70.3	75.0
	BoVW	10^{-3}	63.4	66.7	10^{-4}	68.6	66.5	1.0	65.8	67.6
BoAW + BoVW	early	10^{-6}	66.2	66.5	10^{-1}	67.6	66.2	10^{-6}	71.5	70.5
	late	–	62.2	65.0	–	65.7	64.5	–	62.9	66.0
BoAW + BoVW + BoNG	early	1.0	74.1	75.0	10^{-2}	**81.0**	73.1	10^{-3}	73.6	75.5
	late	–	70.0	79.3	–	70.0	75.0	–	75.7	75.5

gued that the training set augmentation with out of domain data can improve text-based sentiment analysis in some circumstances; similar results are reported in [12].

For the BoAWs, we gained a development set UAR of 69.5 % (Cs 500, Na 1) in the MOVIE REVIEW ONLY set-up (cf. Table 3). For both the CROSS DOMAIN #1 and CROSS DOMAIN #2 set-ups, we achieved small gains in development set accuracy; 71.0 % (Cs 500, Na 10) and 70.3 % (Cs 4000, Na 10) respectively. All set-ups achieved a very similar test set score with the CROSS DOMAIN #2 set-up gaining the highest UAR of 75.0 %. This result indicates that for the BoAW feature small gains can be found with the addition of the extra training data.

The BoVWs representations performed the weakest of the representations tested (cf. Table 3), gaining a UAR for the *Movie Review Only* set-up of 63.4 % (Cs 8000, Na 25). However, gains in system performance were seen in both cross domain set-ups where development UARs of 68.6 % (Cs 8000, Na 25) and 65.8 % (Cs 8000, Na 25) were achieved for the CROSS DOMAIN #1 and CROSS DOMAIN #2 set-ups respectively. As in the BoAWs, the highest test set UAR, 67.6 %, was achieved with the CROSS DOMAIN #2 set-up, highlighting the benefits of adding the extra training samples.

When performing both early and late fusion of the strongest performing BoAW and BoVW systems, there is a slight drop in accuracy when compared to the BoAW only results (cf. Table 3). Interestingly, the strongest results for these tests were again observed in the CROSS DOMAIN #2 set-up. The addition of the BoNG improved the performance; however, we were unable to find a fusion set-up that outperformed the individual BoNG results.

We speculate the strong performance of the BoNG features is due to our use of manually transcribed data in the movie review data, which would be less affected by noise than the other feature spaces. Potentially, these results could be weaker if using automatically transcribed speech which is inherently more susceptible to the effects of noise.

When comparing test set results for the two cross-domain scenarios, a general trend of stronger results can be observed in the CROSS DOMAIN #2 set-up (cf. Table 3). The inclusion of the additional data when forming the dictionary (cf. Section 4), creates bag-of-word feature representations that takes into account aspects of this dataset when quantising the feature spaces. We speculate that this inclusion increases system robustness, although further tests with different datasets from a variety of domains would be needed to verify this.

6. CONCLUSION

This work explored the effects of including cross-domain training data when performing polarity detection sentiment analysis. Our analysis indicates that when forming a bag-of-words feature representation, using a paradigm that takes into account information from both the test and additional datasets can results in gains in system performance. To the best of the authors knowledge this is the first time such a result has been reported for multimodal sentiment analysis.

Future work will focus on verifying the presented finding. We plan to collect further multimodal data sets available on social media from a wide range of spoken reviews and vlogs with an emphasis of collecting data from different cultures. We also plan to utilise multimodal deep unsupervised representation learning methods [30] for sentiment analysis. Finally, we want to explore the effects of including cross domain data when performing analysis of sentiments collected in human-human interactions.

7. ACKNOWLEDGEMENTS

This work was supported by the European Unions's Seventh Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu) and No. 688835 (RIA DE-ENIGMA).

8. REFERENCES

[1] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016.

[2] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093 – 1113, 2014.

[3] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S-F. Chang, and M. Pantic, "A Survey of Multimodal Sentiment Analysis," *Image and Vision Computing*, vol. 35, 2017, 23 pages.

[4] A. Ceron, L. Curini, S. M Iacus, and G. Porro, "Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France," *New Media & Society*, vol. 16, no. 2, pp. 340–358, 2014.

[5] A. Oksanen, D. Garcia, A. Sirola, M. Näsi, M. Kaakinen, T. Keipi, and P. Räsänen, "Pro-Anorexia and Anti-Pro-Anorexia Videos on YouTube: Sentiment Analysis of User Responses," *Journal of Medical Internet Research*, vol. 17, no. 11, pp. e256, 2015.

[6] C. Oh and S. Kumar, "How Trump won: The Role of Social Media Sentiment in Political Elections," in *Proc. of PACIS '17*, Langkawi Island, MY, 2017, p. 48.

[7] L-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proc. of ICMI '11*. 2011, pp. 169–176, ACM.

[8] S. Poria, E. Cambria, N. Howard, G-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50 – 59, 2016.

[9] V. Pérez Rosas, R. Mihalcea, and L. P. Morency, "Multimodal Sentiment Analysis of Spanish Online Videos," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 38–45, May 2013.

[10] L. Teijeiro-Mosquera, J. I. Biel, J. L. Alba-Castro, and D. Gatica-Perez, "What your face vlogs about: Expressions of emotion and big-five traits impressions in youtube," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 193–205, 2015.

[11] D. Sanchez-Cortes, S. Kumano, K. Otsuka, and D. Gatica-Perez, "In the mood for vlog: Multimodal inference in conversational social video," *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 2, pp. 1–24, Jun 2015.

[12] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, Co. Sun, K. Sagae, and L-P. Morency, "YouTube Movie Reviews: Sentiment Analysis in an Audiovisual Context," *IEEE Intelligent Systems Magazine*, vol. 28, no. 3, pp. 46–53, 2013.

[13] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. Schuller, "An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech," in *Proc. of ACM Multimedia '17*, Mountain View, CA, 2017, pp. 478–484, ACM.

[14] S. Amiriparian, N. Cummins, S. Ottl, M. Gerczuk, and B. Schuller, "Sentiment Analysis Using Image-based Deep Spectrum Features," in *Proc. of WASA '17*, San Antionio, TX, 2017, AAAC, IEEE, 6 pages.

[15] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, June 2015.

[16] D. Bollegala, D. Weir, and J. Carroll, "Cross-domain sentiment classification using a sentiment sensitive thesaurus," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 8, pp. 1719–1731, Aug 2013.

[17] D. Bollegala, T. Mu, and J. Y. Goulermas, "Cross-domain sentiment classification using sentiment sensitive embeddings," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 2, pp. 398–410, Feb 2016.

[18] M. Schmitt and B. Schuller, "openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," *Journal of Machine Learning Research*, vol. 18, 2017, 5 pages.

[19] M. Schmitt, F. Ringeval, and B. Schuller, "At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech," in *Proc. of INTERSPEECH '16*, San Francisco, CA, 2016, pp. 495–499, ISCA.

[20] B. Schuller, S. Steidl, A. Batliner, et al., "The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring," in *Proc. of INTERSPEECH '17*, Stockholm, SE, 2017, pp. 3442–3446, ISCA.

[21] S. Amiriparian, S. Pugachevskiy, N. Cummins, S. Hantke, J. Pohjalainen, G. Keren, and B. Schuller, "CAST a database: Rapid targeted large-scale big data acquisition via small-world modelling of social media platforms," in *Proc. of ACII '17*, San Antionio, TX, 2017, AAAC, IEEE, 6 pages.

[22] E. Cambria, Y. Song, H. Wang, and N. Howard, "Semantic multidimensional scaling for open-domain sentiment analysis," *IEEE Intelligent Systems*, vol. 29, no. 2, pp. 44–51, Mar 2014.

[23] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 12, pp. 1–135, 2008.

[24] S.Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, and B.Schuller, "Snore Sound Classification Using Image-based Deep Spectrum Features," in *Proc. of INTERSPEECH '17*, Stockholm, SE, 2017, ISCA, 5 pages.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," in *Proc. of NIPS '12*, Lake Tahoe, NV, 2012, pp. 1097–1105, NIPS.

[26] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. of ACM Multimedia '14*, Orlando, FL, 2014, pp. 675–678, ACM.

[27] T. Baltrušaitis and P. Robinson and L-P. Morency, "Openface: An open source facial behavior analysis toolkit," Lake Placid, NY, 2016, pp. 1–10, IEEE.

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[29] R-E. Fan, K-W. Chang, C-J. Hsieh, X-R. Wang, and C-J. Lin, "LIBLINEAR: A library for Large Linear Classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[30] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Sequence to sequence autoencoders for unsupervised representation learning from audio," in *Proc. of DCASE Workshop*, Munich, Germany, 2017, pp. 17–21, IEEE.