King's Research Portal

Citation for published version (APA):
Mirheidari, B., Bittar, A., Cummins, N., Downs, J., Fisher, H. L., & Christensen, H. (2022). *Automatic Detection of Expressed Emotion from Five-Minute Speech Samples: Challenges and Opportunities.*

Automatic Detection of Expressed Emotion from Five-Minute Speech Samples: Challenges and Opportunities

Bahman Mirheidari[1*], *André Bittar*[2*], *Nicholas Cummins*[2], *Johnny Downs*[2],
Helen L. Fisher[2], *Heidi Christensen*[1]

[*]Joint First Authors
[1]Department of Computer Science, University of Sheffield, UK
[2]Institute of Psychiatry, Psychology & Neuroscience (IoPPN), King's College London, London, UK
b.mirheidari@sheffield.ac.uk, andre.bittar@kcl.ac.uk

Abstract

We present a novel feasibility study on the automatic recognition of Expressed Emotion (EE), a family environment concept based on caregivers speaking freely about their relative/family member. We describe an automated approach for determining the *degree of warmth*, a key component of EE, from acoustic and text features acquired from a sample of 37 recorded interviews. These recordings, collected over 20 years ago, are derived from a nationally representative birth cohort of 2,232 British twin children and were manually coded for EE. We outline the core steps of extracting usable information from recordings with highly variable audio quality and assess the efficacy of four machine learning approaches trained with different combinations of acoustic and text features. Despite the challenges of working with this legacy data, we demonstrated that the degree of warmth can be predicted with an F_1-score of **61.5%**. In this paper, we summarise our learning and provide recommendations for future work using real-world speech samples.

Index Terms: Expressed Emotion, Computational Paralinguistics, Natural Language Processing, Real-world Data, Five-Minute Speech Sample

1. Introduction

Speech-based emotion recognition (SER) technologies and systems have been steadily increasing in prominence in the speech processing literature over the last two decades [1, 2]. Typical SER approaches focus on one of two tasks: (i) the recognition of *discrete* emotions, typically the six 'basic' emotions identified by Ekman [3]; and (ii) *continuous* emotion recognition along a dimensional representation, typically arousal and valence [4]. SER applications can provide feedback within supportive technologies within healthcare systems [5, 6]. For example, SER influenced applications have been proposed in early-diagnosis settings. Recently, speech markers have been used to assist in the inference of *Attachment Condition* in school-age children [7]. This approach used an SER-style approach to recognise if children were emotionally secure or insecure. Similarly, the tracking of emotional engagement can be used to assess the negative impact of dementia on communication [8].

The work presented herein is also based around the concept of automatic speech-based emotion recognition within a clinical application. Specifically, we present the analysis of *Expressed Emotion* (EE), a family environment concept based on caregivers speaking freely about the relative/family member in their care [9, 10]. We focus on the automatic recognition of EE ratings from caregiver speech samples talking about their 5-year-old twins. To the best of the authors' knowledge, this is the first time an automated approach for determining EE in such settings has been attempted. Adding to the challenges of this work is the quality of the audio recordings: they were originally made 20 years ago on audio cassette tapes. Despite the associated challenges, we are able to demonstrate that *degree of warmth*, a component of EE, can be predicted with 61.5% F_1-score.

The rest of the paper is organised as follows: Section 2 presents the relevant background from the fields of psychiatry and psychology, Section 3 describes the set-up, data, and methods of this study, Section 4 contains the results, and concluding remarks are provided in Section 5.

2. Background

In the field of psychiatry, EE refers to the attitudes of caregivers towards their child and comprises criticism, hostility, and/or emotional over-involvement, as well as the degree of warmth shown. For over five decades, levels of EE within families have been studied by psychologists and psychiatrists to determine which adults with mental illness are likely to have the poorest outcomes [11, 12, 13]. EE was originally measured through indepth face-to-face interviews but, due to time constraints, has subsequently been assessed through brief samples of caregivers speaking freely about their child. These interactions are known as the *Five-Minute Speech Sample* (FMSS) [14].

Coding of EE from these easy-to-collect speech samples focuses on the emotions that are apparent when the caregiver speaks about their child, drawing on both what is said and the tone of voice. This coding can contain clinically useful information. For example, EE rated from maternal speech samples plays a causal role in the development of antisocial behavioural problems in children[15] and subsequent serious mental illnesses.[16] Other studies have shown that ratings of negative emotions from parents' speech predict the onset and course of other mental health problems in children, including anxiety, depression, and attention-deficit hyperactivity disorder[17] underlining its usefulness as an early predictor of youth mental health difficulties. However, this promising prediction method is rarely used; the coding of speech is labour-intensive and requires highly trained raters. Moreover, human rating potentially has limited reproducibility as it can be prone to drift and unconscious biases. Automating the assessment of EE could dramatically impact clinical practice, and provide clinicians with an important guide to the likelihood that a young person will develop mental health problems and permit them to effectively target preventive interventions and reduce incidence rates of mental disorders.

arXiv:2203.17242v1 [cs.SD] 30 Mar 2022

3. Experimental setup

This section describes the data used within our presented analysis. We introduce the original cohort study, describe the key steps required to produce the analyzable data and outline the methods used in our experiments.

3.1. Cohort Study

The Environmental Risk (E-Risk) Longitudinal Twin Study tracks the development of a nationally representative birth cohort of 2,232 British twin children born in England and Wales in 1994-1995. They have been comprehensively assessed during home visits at ages 5, 7, 10, 12 and 18 years (with 93% retention). The Joint South London and Maudsley and Institute of Psychiatry Research Ethics Committee approved each phase of the study. When the children were 5 years old, speech samples of approximately five minutes were audio-recorded from caregivers (almost exclusively mothers) in their homes to elicit expressed emotion about each child. Trained interviewers asked caregivers to describe each of their children ("For the next 5 minutes, I would like you to describe [child] to me; what is [child] like?"). The caregiver was encouraged to talk freely but if s/he found this difficult, a series of semi-structured probes were used (e.g., "In what ways would you like [child] to be different?"). These speech samples were coded by two trained raters according to manualised guidelines with high inter-rater reliability (r=0.84–0.90). Ratings included the degree of dissatisfaction/negativity and the degree of warmth that the caregiver expressed towards each child (0=none to 5=high).

3.2. Data

The interviews from the E-Risk study were recorded on cassette tapes using consumer-grade equipment available at the time. The cassettes were maintained in storage for 20 years and may have degraded during this time. The audio quality is highly variable with frequent inaudible passages and white noise. The tapes required digitisation, which was carried out by an external contractor using professional equipment. Although the interviews follow a loose structure based on the use of standard prompt questions, they contain passages of overlapping speech. Furthermore, there is often background chatter, interruptions by young children, and other ambient noises. To enable an analysis of linguistic content and for the training of an automatic speech recogniser (ASR), the interviews had to be transcribed. The significant cost of professional transcription and the small project budget imposed limitations on the amount of material we were able to transcribe. Furthermore, the low quality of the audio complicated the process, and the resulting transcripts contained numerous alignment errors, inaccurate segmentation, incorrect time-stamping, missing passages and some incorrectly rendered words. The final sample contained 37 transcribed interviews coded by human raters for EE, a small proportion of the total.

3.3. Methods

Our experiment aimed at to classify interview samples for the level of warmth expressed by mothers towards their twins. We sought to assess and compare the efficacy of using different combinations of features – acoustic-only, text-only, and both of these together.

We adopted a strategy to increase the amount of available data. Using the Audacity[1] audio editor, we aligned the audio
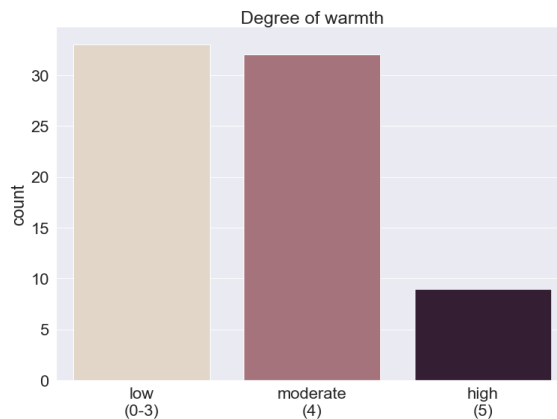
---

[1] https://www.audacityteam.org/

Figure 1: *Distribution of caregiver warmth classes in 3-way schema with corresponding 6-way schema numerical classes.*

and transcripts and manually tagged all utterances to indicate the speaker, which twin was being referred to, and the content type of the utterance, following the loose interview structure. For example, the tag `int-both-support` indicates the interviewer asking about the level of support the mother received during pregnancy, relating to both twins. The tag `mum-t1-away` indicates a mother's description of her feelings when the elder twin (t1) is away from her. The equivalent tag for the younger twin is `mum-t2-away`. Our tagging schema included 38 distinct tags (19 interviewer prompt tags and 19 mother utterance tags). This tagging enabled us to double the size of the dataset by splitting each audio sample into utterances for elder and younger twins. Thus, the dataset used for our experiments contained 74 samples. The EE ratings for each twin were used as target labels for classification.

The original human-based coding of warmth from the E-Risk study consisted of 6 ordinal classes ranging from 0 to 5. However, the distribution of classes was imbalanced in the available dataset. We therefore merged classes into a 3-class schema to provide a more balanced distribution of classes for training our models. The distribution of classes in the final dataset is shown in Figure 1. The resulting distribution, while not perfectly balanced, was more even than the original coding.

4. Results

We used four different machine learning classifiers (with default parameters except for increasing the number of iterations) to predict the degree of warmth for the individual twins: *Logistic Regression* (LR), *Linear Support Vector Classifier* (Lin-SVC), *Random Forest* (RF), and *K-Nearest Neighbours* (KNN). We used the Scikit-learn [18] library in Python to train the classifiers. To evaluate the performance of the classifiers, we ran classification tasks five times using stratified 5-fold cross-validation with shuffling. The final metric was the average of the F_1-score of the classifiers over the runs.

4.1. Acoustic features

We extracted different frame-based acoustic features from the caregivers' audio segments using the OPENSMILE Toolkit [19]. We took the average and standard deviation of the features to build fixed-length feature sets for training. These features were introduced by the authors of OPEN-SMILE at various Interspeech challenges on the detection of

Table 1: *Average F_1-score and standard deviation (5 runs, 5-fold cross validation) of the four classifiers using different acoustic-only features.*

Features	LR	Lin-SVC	RF	KNN
AVEC13	51.6(3.5)	40.2(4.8)	**58.8(2.9)**	39.1(3.0)
IS09-EMO	54.7(2.3)	52.6(2.4)	**57.6(2.5)**	45.3(2.9)
IS10-CMP	59.2(2.1)	59.2(1.7)	**59.9(2.2)**	55.2(2.6)
IS13-CMP	52.3(3.6)	43.1(3.9)	**57.1(1.5)**	38.9(2.9)
IS16-CMP	49.5(3.4)	43.1(3.9)	**56.6(2.0)**	38.9(2.9)
eGeMAPS	**59.9(4.4)**	47.7(7.4)	58.3(2.1)	42.4(2.8)
-	54.5(4.2)	47.6(7.1)	**58.0(1.2)**	43.3(6.4)

The best result in a row is in bold.

Table 2: *Average F_1-score and standard deviation (5 runs, 5-fold cross validation) of the four classifiers using different text-only features.*

Features	LR	Lin-SVC	RF	KNN
TF-IDF	**43.4(3.2)**	41.4(2.7)	42.9(5.4)	43.2(4.2)
TF-IDF-PUN	39.9(3.6)	42.4(3.9)	**46.6(2.9)**	37.5(2.3)
BERT	37.9(2.6)	37.0(2.3)	36.5(2.8)	**41.4(3.7)**
BERT-PUN	49.2(2.9)	**53.4(3.9)**	46.4(4.7)	44.0(3.3)
roBERTa	43.6(3.5)	43.4(6.1)	**45.9(2.9)**	41.1(1.8)
roBERTa-PUN	49.9(3.3)	**53.7(2.7)**	42.3(4.2)	48.1(3.7)
Word2Vec	44.4(3.9)	47.5(3.1)	47.3(2.7)	**47.8(4.1)**
FastText	38.7(5.1)	42.9(4.8)	**46.7(5.5)**	40.6(2.8)
GloVe	43.4(3.9)	48.4(3.8)	53.5(3.5)	43.5(2.2)
-	43.4(4.2)	**45.6(5.6)**	45.3(5.6)	43.0(3.4)

The best result in a row is in bold.

Table 3: *Average F_1-score and standard deviation (5 runs, 5-fold cross validation) of the four classifiers using IS10-CMP acoustic features combined with different text features (Comb. Feat.).*

Comb. Feat.	LR	Lin-SVC	RF	KNN
TF-IDF	59.2(2.1)	**61.0(3.2)**	53.6(3.3)	54.6(2.0)
TF-IDF-PUN	59.0(2.2)	**61.5(2.3)**	52.8(2.7)	54.3(2.1)
BERT-PUN	**59.8(2.6)**	58.4(2.8)	45.3(4.3)	55.4(3.3)
roBERTa	59.9(2.4)	**60.4(2.6)**	46.2(2.1)	55.1(2.1)
roBERTa-PUN	**58.8(2.3)**	58.0(2.8)	45.9(4.5)	55.4(2.5)
Word2Vec	59.2(2.1)	**59.3(2.5)**	52.9(1.7)	54.7(2.3)
FastText	59.2(2.1)	**60.7(2.4)**	52.9(4.9)	53.5(2.5)
GloVe	59.5(2.2)	**60.8(4.1)**	57.1(1.5)	53.8(2.8)
-	59.3(0.4)	**60.0(1.3)**	50.8(4.4)	54.6(0.7)

The best result in a row is in bold.

emotion and paralinguistic characteristics of speech, including the *Interspeech 2009 Emotion Detection Challenge* (IS09-EMO) [20], *Audio-Visual Emotion recognition Challenge 2013* (AVEC13) [21], Interspeech 2010, 2013 and 2016 *Computational Paralinguistics ChallengE* (ComParE) (IS10-CMP, IS13-CMP, IS16-CMP) [22, 23, 24], and the *extended Geneva Minimal Acoustic Parameter Set* (eGeMAPS) [25].

Table 1 shows the average F_1-score and standard deviation of the four classifiers on different acoustic features. The results varied across combinations of features and classifiers. The RF classifier using IS-10 features and the LR classifier with eGeMAPS resulted in the best average F_1-score of 59.9%; the former having fewer errors (2.2% vs. 4.4%). KNN on both IS13-CMP and IS16-CMP scored lowest (F_1=38.9%). IS13-CMP and IS16-CMP features on all classifiers compared to IS10-CMP features had statistically significantly lower F_1-scores (p-value < 0.05). In addition, the results obtained by the RF classifier were significantly better than the other classifiers.

4.2. Text features

We tested three methods to represent the textual content of the interviews: i) Term Frequency-Inverse document frequency using the Scikit-Learn [18] `TfIdfVectorizer` with default settings fitted to the training set (TF-IDF); ii) pre-trained word emeddings: GloVe [26] 25 dimensions, trained on Twitter data (GloVe), FastText [27] 300 dimensions, trained on Wikipedia (FastText), Word2Vec [28] 300 dimensions, trained on the Google News corpus (Word2Vec); iii) Pre-trained transformer-based language models[2]: BERT [29] base, large, uncased model (BERT), and roBERTa [30] large base uncased model (roBERTa).

We preprocessed the transcripts using spaCy [31], tokenising the text, removing punctuation and lingering whitespace, and lowercasing all tokens. For the pre-trained language models, which are limited to sequences of 512 tokens, the text was divided into chunks of 512 tokens and passed to the models using a sliding window approach with 50% overlap and the average and standard deviation of the last three layers was computed to create features for classification. For TF-IDFand pre-trained language models, we trialled using the original transcripts with punctuation (reported in results with the suffix *"-pun"*) and the pre-processed texts. Word embedding models used the pre-processed text only. For the embedding models, unknown words (words that did not appear in the model) were ignored. The final embedding representation for each transcript was the mean of embeddings for all known word tokens in the

transcript. In all runs, only utterances for caregivers were retained as this was the content used for the manual coding of warmth in the transcripts. We used the same classification models for the acoustic features.

Table 2 shows the average F_1-score and standard deviation of the four classifiers on the different text features. The best average F_1-score was achieved by a Lin-SVC classifier using the roberta-pun features (F_1=53.7%), while the lowest performance was obtained with RF and bert features (F_1=36.5%). The bert features on all classifiers also had significantly lower F_1-scores (p-value < 0.05) compared to roberta-pun features. However, the difference between the classifiers' performance was not statistically significant.

4.3. Combining acoustic and text features

The manual coding of warmth (and EE in general) relies on both interview content and voice features. We therefore sought to assess the use of both modalities in the classification task, using a combination of acoustic and text features to train the models. Since the IS10-CMP acoustic features yielded the best results on all classifiers, we combined these features with each of the text features, with the exception of BERT, due to its significantly lower performance in the text-only task. The average F_1-score of the classifiers on the combined features are shown in Table 3[3]. The overall figures across different clas-

---

[3]Note that we tried to normalise the features before the combination, however we could not see much difference in the results, so we ignored
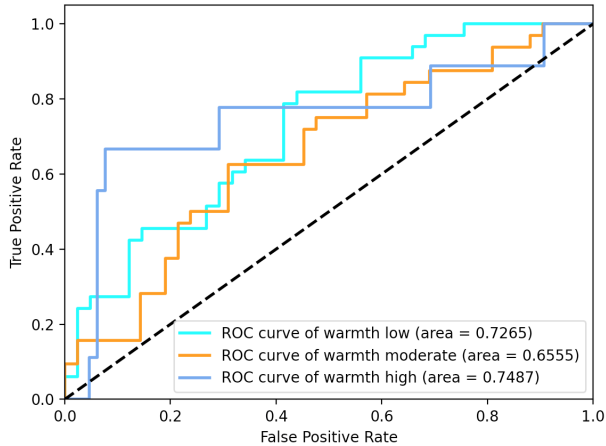
Figure 2: *Receiver operating characteristic curves of the three-way Lin-SVC classification using is10-par acoustic features combined with TF-IDF-PUNtext features.*
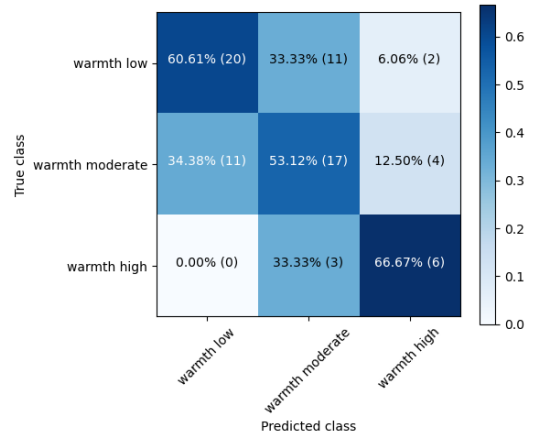

Figure 3: *Confusion matrix of the three-way Lin-SVC classification using is10-par acoustic features combined with TF-IDF-PUNtext features.*

sifiers and features showed an increase in performance compared to both acoustic-only and text-only experiments. The best performance was obtained by the Lin-SVC classifier using TF-IDF-PUNfeatures (F_1=61.5%). The second best performance was obtained by the same classifier using TF-IDFfeatures on the pre-processed text (F_1=61%). A t-test showed no significant differences between the F_1-scores across all features. However, Lin-SVC and LR classifiers performed significantly better than KNN and RF.

Figure 2 shows the Receiver Operating Characteristic (ROC) curves for one of the representative runs of the best performing classifier (Lin-SVC) on IS10-CMP combined with TF-IDF-PUNfeatures. Figure 3 shows the corresponding Confusion Matrix (CM). The best ROC curve (up to 0.7487) was obtained for *high* warmth (the minority class). Highest accuracy was also achieved for this class (67% or 6 out of 9), followed by the *low* class (the majority class) with accuracy of around 61% (20/33). The most challenging class to predict was *moderate* which had the highest degree of confusion (with *low* warmth) and an accuracy of 53%.

5. Conclusions

This study has highlighted some of the significant challenges of working with a dataset of *real-world* audio. First, the low quality of audio recorded on cassette tapes and low-grade equipment greatly complicated the task of transcription and significantly increased the investment required to make the dataset usable. Second, the cost of transcription and a restricted project budget limited the number of transcripts we could obtain. Manually tagging the utterances in the interviews proved an effective way of increasing the amount of data and enabled us to split interviews into two parts, one for each twin. Third, the small amount of available data imposed limitations on the choice of classification models we could use. It was not feasible to trial state-of-the-art models that require much larger amounts of data, such as deep neural networks. Finally, due to class imbalance, it was necessary to merge the classes of our target variable. Our experiments should, therefore, be seen as a proxy for the task of predicting true warmth scores as coded in the E-Risk study.

_____

the normalisation when combining features.

Despite these limitations, our results tentatively indicate that combining acoustic and text features is optimal when trying to predict the levels of caregiver warmth expressed in Five Minute Speech Samples. This promising result suggests that machine learning classifiers may eventually be an adequate substitute for the process of manual coding of warmth, and EE more generally, by human raters.

In future work, we intend to prioritise the expansion of the dataset with additional transcriptions. A larger dataset would open up the possibility of using more sophisticated classification models. Ultimately, however, we aim to develop an approach based on automatic speech recognition in order to alleviate the burden of manual transcription.

We conclude by making recommendations for researchers faced with the challenges of working with *real-world* speech samples. Investing time in the manual preparation of data, such as tagging or other annotations, can help mitigate the effects of limited and low quality data. We suggest adapting experiments to the limitations of the data by using a variety of established and recent feature extraction and machine learning methods. Finally, in initial experiments consider using a combination of a reduced number of classes and conventional machine learning approaches, as this should help keep model complexity lower when only smaller amounts of data are available.

6. Acknowledgements

7. References

[1] B. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90—-99, 2018.

[2] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 93–120, 2018.

[3] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[4] P. Kuppens, F. Tuerlinckx, J. A. Russell, and L. F. Barrett, "The relation between valence and arousal in subjective experience," *Psychological Bulletin*, vol. 139, no. 4, p. 917, 2013.

[5] J. M. Harley, S. P. Lajoie, C. Frasson, and N. C. Hall, "Developing Emotion-Aware, Advanced Learning Technologies: A Taxonomy of Approaches and Features," *International Journal of Artificial Intelligence in Education*, vol. 27, no. 2, pp. 268–297, 2017.

[6] A. Jaimes and N. Sebe, "Multimodal human-computer interaction: A survey," *Computer Vision and Image Understanding*, vol. 108, no. 1, pp. 116–134, 2007.

[7] H. Alsofyani and A. Vinciarelli, "Stacked Recurrent Neural Networks for Speech-Based Inference of Attachment Condition in School Age Children," in *Proc. Interspeech 2021*. Brno, Czechia: ISCA, 2021, pp. 2491–2495.

[8] L. Steinert, F. Putze, D. Küster, and T. Schultz, "Audio-Visual Recognition of Emotional Engagement of People with Dementia," in *Proc. Interspeech 2021*. Brno, Czechia: ISCA, 2021, pp. 1024–1028.

[9] J. Kanter, H. R. Lamb, and C. Loeper, "Expressed emotion in families: A critical review," *Psychiatric Services*, vol. 38, no. 4, pp. 374–380, 1987.

[10] E. D. Hibbs, S. D. Hamburger, M. Lenane, J. L. Rapoport, M. J. Kruesi, C. S. Keysor, and M. J. Goldstein, "Determinants of expressed emotion in families of disturbed and normal children," *Journal of Child Psychology and Psychiatry*, vol. 32, no. 5, pp. 757–770, 1991.

[11] G. W. Brown, J. L. Birley, and J. K. Wing, "Influence of family life on the course of schizophrenic disorders: a replication," *The British Journal of Psychiatry*, vol. 121, no. 562, pp. 241–258, Sep. 1972.

[12] M. Rutter and G. W. Brown, "The reliability and validity of measures of family life and relationships in families containing a psychiatric patient," *Social psychiatry*, vol. 1, pp. 38–53, Aug. 1966.

[13] C. E. Vaughn and J. Leff, "The measurement of expressed emotion in the families of psychiatric patients." *British Journal of Social and Clinical Psychology*, vol. 15, no. 2, pp. 157–65, 1976.

[14] A. B. Magaña, J. M. Goldstein, M. Karno, D. J. Miklowitz, J. Jenkins, and I. R. Falloon, "A brief method for assessing expressed emotion in relatives of psychiatric patients," *Psychiatry Research*, vol. 17, no. 3, pp. 203–212, Mar. 1986.

[15] A. Caspi, T. E. Moffitt, J. Morgan, M. Rutter, A. Taylor, L. Arseneault, L. Tully, C. Jacobs, J. Kim-Cohen, and M. Polo-Tomas, "Maternal expressed emotion predicts children's antisocial behavior problems: using monozygotic-twin differences to identify environmental effects on behavioral development," *Developmental Psychology*, vol. 40, no. 2, pp. 149–161, Mar. 2004.

[16] T. E. Moffitt, A. Caspi, H. Harrington, and B. J. Milne, "Males on the life-course-persistent and adolescence-limited antisocial pathways: follow-up at age 26 years," *Development and Psychopathology*, vol. 14, no. 1, pp. 179–207, 2002.

[17] E. Sher-Censor, "Five Minute Speech Sample in developmental research: A review," *Developmental Review*, vol. 36, pp. 127–155, 2015.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[19] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*. Firenze, Italy: ACM, 2010, pp. 1459–1462.

[20] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 Emotion Challenge," in *Proc. Interspeech*. Brighton, Uk: ISCA, 2009, pp. 312–315.

[21] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013 – The Continuous Audio/Visual Emotion and Depression Recognition Challenge ," in *Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge*. Barcelona, Spain: ACM, 2013, pp. 3–10.

[22] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The Interspeech 2010 Paralinguistic Challenge," in *Proc. Interspeech*. Makuhari, Japan: ISCA, 2010, pp. 2794–2797.

[23] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The Interspeech 2013 Computational Paralinguistics Challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech*. Lyon, France: ISCA, 2013, pp. 148–152.

[24] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini *et al.*, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proc. Interspeech*. San Francisco, CA, USA: ISCA, 2016, pp. 2001–2005.

[25] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[26] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*. Lisbon, Portugal: ACL, 2014, pp. 1532–1543.

[27] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," 2016. [Online]. Available: https://arxiv.org/abs/1607.01759

[28] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., vol. 26. Curran Associates, Inc., 2013.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018. [Online]. Available: https://arxiv.org/abs/1810.04805

[30] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[31] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python," 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3358113,