



Is Everything Fine, Grandma? Acoustic and Linguistic Modeling for Robust Elderly Speech Emotion Recognition

Gizem Soğancıoğlu¹, Oxana Verkholyak², Heysem Kaya¹, Dmitrii Fedotov^{3,4}, Tobias Cadée¹, Albert Ali Salah¹, Alexey Karpov²

¹ Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands

² St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences
SPIIRAS, St. Petersburg, Russia

³ Ulm University, Ulm, Germany

⁴ ITMO University, St. Petersburg, Russia

g.sogancioglu@uu.nl, overkholyak@gmail.com, h.kaya@uu.nl, dmitrii.fedotov@uni-ulm.de, cadeetobias@gmail.com, a.a.salah@uu.nl, karpov@iias.spb.su

Abstract

Acoustic and linguistic analysis for elderly emotion recognition is an under-studied and challenging research direction, but essential for the creation of digital assistants for the elderly, as well as unobtrusive telemonitoring of elderly in their residences for mental healthcare purposes. This paper presents our contribution to the INTER_SPEECH 2020 Computational Paralinguistics Challenge (ComParE) - Elderly Emotion Sub-Challenge, which is comprised of two ternary classification tasks for arousal and valence recognition. We propose a bi-modal framework, where these tasks are modeled using state-of-the-art acoustic and linguistic features, respectively. In this study, we demonstrate that exploiting task-specific dictionaries and resources can boost the performance of linguistic models, when the amount of labeled data is small. Observing a high mismatch between development and test set performances of various models, we also propose alternative training and decision fusion strategies to better estimate and improve the generalization performance.

Index Terms: speech emotion recognition, human-computer interaction, computational paralinguistics, sentiment analysis

1. Introduction

While the state-of-the-art in affective computing and paralinguistic analysis reaches new peaks, research on two subject groups, namely the children and the elderly, lags behind due to scarcity of training resources and difficulty of data collection [1, 2]. Acoustic characteristics of these groups differ significantly from other age groups mainly found in the available datasets. Models trained on available data do not perform optimally for recognition in extreme age groups. Computational Paralinguistics Challenge (ComParE) 2020 [3, 4] introduces a novel elderly emotion dataset, where both acoustic signals and speech transcriptions are provided.

Recent ComParE challenges introduced new feature types, such as Bag-of-Audio-Words [5] and embeddings from Sequence-to-Sequence Deep Recurrent Neural Networks [6]. The results of these challenges showed that ensemble systems and alternative feature representations have a great potential for advancing the state-of-the-art. However, the constituents and the combination rules of the ensemble systems must be selected with care. Kaya and colleagues previously applied the Fisher Vector (FV) encoding of acoustic Low-Level Descriptors (LLD)

to several paralinguistic tasks including recognition of native language and sincerity [7], as well as classification of snoring types [8] and eating conditions [9]. In this work, we use a similar FV encoding for the representation of acoustic features.

In affective computing and paralinguistics research, it is known that acoustic models perform well for arousal recognition, while providing poorer performance on valence recognition [10, 11, 12]. Leveraging video and spoken content, when they are available, provides significant improvement on both valence and categorical emotion recognition performance [12]. Based on our experiences and the relevant literature [3, 10, 11], we propose to leverage the linguistic modality for valence and the acoustic modality for arousal in this work. We propose modality-specific ensemble systems for arousal and valence recognition, while investigating the effectiveness of acoustic and linguistic models on both recognition tasks to support our hypothesis. For robust valence modeling, we extract a set of state-of-the-art linguistic features, including TF-IDF (Term Frequency-Inverse Document Frequency), FastText word embeddings, high-level polarity features, and dictionary-based linguistic features in German and English.

The contribution of this work is manifold. Firstly, we propose a bi-modal framework leveraging the linguistic modality for valence and the acoustic modality for arousal prediction. Secondly, we extract and experiment with a plethora of state-of-the-art acoustic and linguistic feature sets. Thirdly, we investigate strategies for fusion (at both feature and decision levels) and modeling with a high generalization power. We apply the proposed systems and strategies on the ComParE-2020 Elderly Emotion Sub-Challenge and obtain a marked improvement over the challenge test set baselines.

2. Background on Methods

Our task in this work is to predict valence and arousal ratings (in Low, Medium, High levels) of spontaneous narratives (i.e. acoustic and linguistic modalities). The reader is referred to [3] for details of the baseline acoustic and linguistic features, as well as the challenge corpora. Here, we briefly provide the background on methods we used from the literature.

2.1. Fisher Vector Representation for Acoustic Descriptors

The Fisher Vector (FV) encoding [13] is a state-of-the-art representation method for representing low level descriptors (LLD)

over an image, utterance or video, firstly introduced in the computer vision domain, and successfully applied in paralinguistic analysis [7, 8]. For the emotion estimation task, we train FVs for the acoustic modality at utterance level. FV requires a background probability model, typically a Gaussian Mixture Model (GMM), trained on LLDs of all the utterances (more on this later) from the training set. Normally, this requires the computation of the Fisher information matrix, which can be approximated in the case when diagonal covariance matrices are used with the GMM. The set of LLDs are de-correlated and projected to a lower dimension using Principal Component Analysis (PCA). Hence, the number of PCA components K_{PCA} and GMM clusters K_{GMM} are the hyper-parameters of the FV encoding.

2.2. Sentiment Dictionaries

We use two sentiment dictionaries to estimate emotion from language use. These are language specific resources that contain a list of affective words with associated positive or negative scores. The SentiWS dictionary [14] contains 3467 German words together with the corresponding inflections and Part-of-Speech (POS) tags. Each word is assigned a single score, which is estimated using frequencies and co-occurrence statistics on a German-language corpus consisting of approximately 100M sentences. The scores are scaled to the range of [-1.0, 1.0] with +1.0 being absolutely positive and -1.0 being absolutely negative. The SentiWordNet dictionary [15] is based on WordNet and consists of almost 207K English words (60 times more than SentiWS) together with POS tags. Each word is assigned a positive and a negative score in the range of [0.0, 1.0].

2.3. Supervised Classifiers

Due to their popularity in handling high-dimensional feature vectors (e.g. supra-segmental acoustic features), Support Vector Machines (SVMs) are used in many emotion classification systems. In our ensemble, we additionally employ Kernel Extreme Learning Machines (ELM) [16] and Partial Least Squares (PLS) regression [17], since these are fast and accurate algorithms that previously produced state-of-the-art results on several speech-based and multimodal tasks [7, 8]. We obtain kernels from the training data for both PLS and ELM. For handling data imbalance, we employ a variant of ELM dubbed *Weighted Kernel ELM (WKELM)* [18] (and its Kernel PLS counterpart WKPLS that we introduced in [8]), which assign higher weights to minority class instances during model learning.

In addition to kernel-based classifiers, we employ Gradient Boosting Machines (GBM) in this work. GBMs are a special family of decision tree ensembles, where the K^{th} tree is trained to predict the residual from the former $K - 1$ trees [19, 20]. In GBMs, tree learning is boosted with instance-wise gradient and the Hessian of the loss function.

3. System Development

An overview of the proposed bi-modal arousal and valence recognition system is given in Figure 1. The main idea is to leverage the strength of different modalities in each task. The details of feature extraction, classification and fusion steps are given in the following subsections.



Figure 1: *The proposed pipeline for bi-modal elderly speech emotion recognition.*

3.1. Acoustic Feature Extraction

We extract FV based features that were shown to be effective in former paralinguistic challenges [7, 8]. As acoustic LLDs, we extract Mel-Frequency Cepstral Coefficients (MFCCs) 0-24 and RASTA-PLP (Perceptual Linear Prediction) cepstrum for 12^{th} order linear prediction, together with their temporal Δ coefficients, making an LLD vector of 76 dimensions. The combined LLDs are PCA-projected to preserve 99.9% of the variation in the data. This keeps the dimensionality virtually unchanged (75 dimensions), but de-correlates the data. The number of GMM clusters are optimized using cross-validation (CV).

3.2. Linguistic Feature Extraction

Transformer language embeddings (such as BERT [21]) are the state-of-the-art in representing linguistic features, and are also a part of the baseline system for the given sub-challenge. However, as seen from the baseline performance, they may show unreliable results on small datasets. Therefore, we test four alternative representations: FastText, dictionary-based, high-level polarity, and TF-IDF features, respectively. We explain these features in dedicated subsections.

In addition to representations obtained from the original German transcriptions, we propose to extract features from the automatic English translations obtained via Google Translation engine¹. Analysing English texts is advantageous, since there are more sentiment analysis resources available for English.

3.2.1. FastText Embeddings

We use the FastText model [22], which is a state-of-the-art approach for character level embeddings, producing a semantic vector representation of words in a story. We use pre-trained 100-dimensional English and German word embeddings [23], which are trained on Common Crawl², and finetune the pre-trained model on our dataset. The story-level vectors are obtained by averaging the 100-dimensional vectors of each word that construct the corresponding story.

3.2.2. Dictionary-based Features

In order to obtain story-level SentiWS scores, an input text is tokenized ignoring the punctuation, and each token is looked up in the dictionary. If it is not found, the list of inflections for each word in the dictionary is checked without POS tag. Usually this results in a single score, but if multiple matches are found, a mean between the scores is accepted as a final score. The output of this process is a sequence of scores for tokens found in the dictionary. The following statistics are applied to the sequence: minimum, maximum, range, mean, sum, and numbers of positive and negative scores.

For the SentiWordNet representation, an input text is tokenized ignoring the punctuation, and each token is looked up ac-

¹<https://cloud.google.com/translate>

²<http://commoncrawl.org/>

ording to its POS. It is common to see multiple matches since English is rich in homographs and SentiWordNet disambiguates between many of them. The mean score is used as a final score. If a token is not found in the dictionary, the same process is repeated for its lemma using original token’s POS. The outputs of this process are two sequences containing positive and negative scores for each token found in the dictionary. Then the sequences are used to calculate two sets of the following statistics: minimum, maximum, range, mean, sum and number of instances. The tokenization, POS tagging and lemmatization were performed using the NLTK Python library [24].

3.2.3. High-level Polarity Features

As polarity and subjectivity features, we use available sentiment analysis tools, namely NLTK Vader [25], TextBlob [26] and Flair [27]. Each of these libraries have some strengths and drawbacks in assessing the sentiment of the sentences. For instance, TextBlob is based on a simple pattern analyzer logic and fails to take negation into account in the sentence. However, alongside polarity prediction, it applies subjectivity analysis, which can be considered as a good feature for the valence dimension. Vader is good at handling negation thanks to some heuristics, but performs weakly on unseen words. Flair, which is based on a character-level Long Short-Term Memory (LSTM) network, is good at assessing the polarity of unseen words (such as those that result from typos).

To benefit from the strengths of each approach, we use the predicted polarity and subjectivity probability scores of each library as high-level features for our model. Since English is the common supported language among these tools, we use English transcripts (machine translated from German) to extract these features. Although these methods are designed to work on sentence-level or shorter length of text, we applied them at the story-level yet still obtain a good performance.

3.2.4. TF-IDF Model

We extract Term Frequency-Inverse Document Frequency (TF-IDF) representation as an additional linguistic feature. This representation is commonly used in natural language processing [28], information retrieval [29] and text mining [30] tasks. As a vital step of using lexical level features, we apply standard pre-processing methods, such as removal of stop words by using the English/German stop words dictionaries available in the NLTK library [24], and stemming by the Porter stemmer [31]. Afterwards, TF-IDF weights are computed over the set of unigrams and bi-grams.

3.3. Proposed Model Training and Optimization Strategy

A typical training strategy under the challenge protocol is as follows. We train the models on the challenge *training set*, optimize the hyper-parameters and the feature types on the *development set*. Once an optimal setting is found, we combine both and re-train with the optimal setting. Generalization to the sequestered test set depends on many factors, and is difficult to predict. To overcome this problem, we propose to use N-Fold cross-validation (CV) to generate N learners and to fuse their decisions for the test set. This effectively increases the training data, shows the model’s performance on the entire set of annotated data points and also reduces the test set error via combining multiple learners. To estimate the real-life (in our case challenge test set) performance, we propose to use a nested N-Fold CV.

3.4. Proposed Label Fusion Strategies

For fusing class labels, it is common to use majority voting of three or more models. We extend this approach, mainly for breaking ties, and we propose a rule based strategy for fusing two prediction sets. The proposed tie break mechanisms benefit from our domain knowledge, namely the ordinal structure of the target variables and the class distribution.

The tie break mechanism for three models in a ternary classification uses an idea that in such a tie all three levels are predicted by base models, thus finds the middle way, outputting the medium class. In the case a higher number (> 3) of models are combined, it favors the minority class(es). The tie break mechanism for two models further considers the development set confusion matrices to infer the bias towards majority and minority classes of the two models. A rule set used for combining a prediction set favoring the majority class (P) and a prediction set favoring the minority classes (S) is given in Algorithm 1.

Algorithm 1 Fusion of two prediction sets $\{P_i, S_i\}$ into $\{O_i\}$, $i=1 \dots N$, where the class labels $\{‘L’, ‘M’, ‘H’\}$ are ordinal.

```

if  $P_i = S_i$  then
     $O_i \leftarrow P_i$ 
else if  $P_i$  and  $S_i$  contain opposite extreme labels then
     $O_i \leftarrow ‘M’$ 
else if  $S_i$  contains a minority class label then
     $O_i \leftarrow S_i$ 
else
     $O_i \leftarrow P_i$ 
end if

```

4. Experimental Results

The Ulm State of Mind Elderly (USOMS-e) database consists of 87 participants (55 f), who reported three narratives each. These are scored for valence and arousal, which are then grouped into three levels. For further details, see [3].

4.1. Experiments with Baseline Feature Sets

Using the conventional train-to-development setting, we obtained the best development set arousal Unweighted Average Recall (UAR) of 44.4% using PCA reduced auDeep-60 features [6] with an SVM classifier that yielded a slightly lower test set UAR score of 42.7%. It should be noted that the best test set arousal performance (UAR=50.4%) is obtained using Deep-Spectrum ResNet50 [32] features, where the development set UAR was 35.0%. Using the baseline feature sets, our best development set valence UAR=56.1% was obtained with the combined BERT [21] feature dubbed ‘BLAtt’ and German POS tags (BLAtt+POS) features modeled with a GBM classifier. This linguistic system yielded a test set score of 42.8%. We attribute this performance mismatch partly to the model optimization/training procedure as discussed in Section 3.3. Hereafter, we follow our pipeline using the proposed features with 4-Fold CV for model training.

4.2. Experiments with FV Representation

We carried out FV feature extraction using a number of GMM components, $K_{GMM} \in \{16, 32, 64, 128\}$. Although there was one annotation for each story, the challenge utterances were provided as chunks of 5 seconds, particularly for acoustic modeling. We hypothesized that summarizing the LLDs over each

story and carrying out direct classification on story-wise FVs would yield a better performance than representing chunks and then carrying out decision fusion over each development/test set story. This hypothesis was tested on the training/development set using $K_{GMM}=64$. We observed that this approach boosted both arousal (from 42.7% to 48.2%) and valence (from 45.8% to 51.0%) UAR performance on the development set. Thus, we conducted remaining experiments using story-wise FVs.

Noting a small number of story-wise instances (87 train + 87 dev = 174), we preferred smaller K_{GMM} if a similar performance is obtained. Using FV representation with 16 GMM components, we obtained a 4-Fold CV overall UAR score of 48.7% (KELM, linear kernel) for arousal and 52.0% for valence (KPLS, linear kernel). We further reduced the dimensionality of FV using PCA and obtained the best UAR performance with 150 and 160 PCA dimensions for arousal (50.1% with KELM, 45.1% with WKELM) and valence (55.3% with KPLS, 53.8% with WKPLS), respectively. The corresponding weighted score fusion gave a 4-Fold CV UAR scores of 51.9% and 58.2% for arousal and valence, respectively. We used this system for our test set submission.

4.3. Experiments with Linguistic Features

Using the experimental setup defined in Section 3.3, we evaluated the performance of different combinations of linguistic features, as well as combinations of the models with hard-label-based majority voting. The linguistic models were trained using SVMs with linear, sigmoid and radial basis function kernels. As can be seen in Table 1, each proposed linguistic approach performed significantly better than the baseline reported in the challenge paper for valence recognition [3]. Moreover, a combination of these models at the decision level, namely the Ensemble Model, contributed to the overall performance. On the other hand, as expected, we obtained poor performance using linguistic features for arousal classification.

Considering the best UAR performances of the baseline system for valence (56% for development and 49% for test) that uses the BERT model [21], and those of our FastText features [22] (4-fold CV UAR: 46.5%), we observed that using only the complex contextual/semantic word embeddings to represent the story may be insufficient when the amount of labeled data is relatively small. But enriching those features with the knowledge learned from some external resources (such as tonal dictionaries or sentiment analysis tools) that are trained on much larger data sets can greatly improve the performance.

Except high-level polarity features, we conducted experiments for all models using both English translations and German transcripts. For FastText and TF-IDF models, using only English text gave a slightly better performance than using bilingual text. Thus, for these features, we reported the models that employ feature extraction from English text. Regarding dictionary based features, although the original text was in German, English dictionary-based features performed better than German ones, while fusion of the two followed by feature selection yielded the best performance. Brute-forcing different feature combinations resulted in the following set of optimal dictionary-based features: 2 from SentiWordNet (maximum positive and sum of negative scores) and 3 from SentiWS (minimum, maximum and number of negative scores). This small set of dictionary-based features outperformed the state-of-the-art BERT features on the development set by an absolute difference of 16.2%.

Table 1: 4-Fold CV Average UAR (%) of Linguistic Models. Ensemble Model: (1, 4, 5) for valence, and (1, 3, 4, 5) for arousal.

Features	Dimens.	Valence	Arousal
(1) TF-IDF features	20337	52.3	33.8
(2) FastText features	100	46.5	31.3
(3) Polarity features	7	57.0	40.4
(4) FastText+Polarity features	107	60.9	36.3
(5) Dictionary-based features	5	61.9	34.4
Ensemble Model	-	62.3	38.1

4.4. Challenge Test Set Submissions

A total of five test set submissions were evaluated to prevent learning on the sequestered test set. A summary of the test performances of the proposed features are given in Table 2. These results are in line with our hypotheses and show that linguistic models have a high generalization ability. System 3 arousal performance on the test set also shows that domain-aware rule based fusion of two prediction sets given in Algorithm 1 dramatically improves the performance. Reaching an average test set UAR performance of 60.6%, we outperform the challenge baseline relatively by 21.9%.

Table 2: Test Set UAR (%) Performances

System	Description	Valence	Arousal
System 0	Challenge baseline	49.0	50.4
System 1	FV based Ensemble	44.3	48.8
System 2	Linguistic Ensemble	63.7	41.2
System 3	Valence: Same as System 2, Arousal: Rule based fusion System 0 + System 1	63.7	57.5

5. Conclusions

The proposed bi-modal elderly emotion recognition system shows outstanding performance on both arousal and valence recognition tasks with mean UAR=60.6%, beating the baseline mean by an absolute difference of 10.9%. Simple linguistic features proposed in this study dramatically outperform state-of-the-art BERT systems both in terms of accuracy and generalization capability that shows their advantage when using a small dataset. Moreover, they provide explainable results in contrast to “black-box” approaches. Linguistic modeling on automatically translated English text shows better performance than the original, because of better resources. Similar results were reported for other languages [33]. Together with N-Fold CV experimental set-up and careful ensemble building strategy, the proposed system allows obtaining high reliability in terms of performance on a blind test set, which is not always possible with a traditional train/development split. Finally, it is shown that domain and confusion matrix awareness proves to be of considerable practical importance at the final stage of decision-level fusion. Scripts of this work can be found at <https://github.com/gizemsogancioglu/elderly-emotion-SC>.

6. Acknowledgements

This research was supported by the Russian Science Foundation (project No. 18-11-00145).

7. References

- [1] K. Wang, Q. Zhang, and S. Liao, "A database of elderly emotional speech," in *Proc. International Symposium on Signal Processing, Biomedical Engineering and Informatics*, 2013, pp. 549–553.
- [2] H. Kaya, A. A. Salah, A. Karpov, O. Frolova, A. Grigorev, and E. Lyakso, "Emotion, age, and gender classification in children's speech by humans and machines," *Computer Speech & Language*, vol. 46, pp. 268 – 283, 2017.
- [3] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stapfen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly emotion, Breathing & Masks," in *INTERSPEECH*, Shanghai, China, October 2020, to appear.
- [4] M. Markitantov, D. Dresvyanskiy, D. Mamontov, H. Kaya, W. Minker, and A. Karpov, "Ensembling end-to-end deep models for computational paralinguistics tasks: ComParE 2020 Mask and Breathing Sub-challenges," in *INTERSPEECH*, Shanghai, China, October 2020, to appear.
- [5] M. Schmitt and B. Schuller, "OpenXBOW: introducing the pasau open-source crossmodal bag-of-words toolkit," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 3370–3374, 2017.
- [6] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "Audeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *The Journal of Machine Learning Research*, vol. 18, no. 1, p. 6340–6344, 2017.
- [7] H. Kaya and A. Karpov, "Fusing acoustic feature representations for computational paralinguistics tasks," in *INTERSPEECH*, San Francisco, USA, 2016, pp. 2046–2050.
- [8] H. Kaya and A. A. Karpov, "Introducing weighted kernel classifiers for handling imbalanced paralinguistic corpora: Snoring, addressee and cold," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3527–3531.
- [9] H. Kaya, A. A. Karpov, and A. A. Salah, "Fisher vectors with cascaded normalization for paralinguistic analysis," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 909–913.
- [10] B. Schuller, "Voice and speech analysis in search of states and traits," in *Computer Analysis of Human Behavior*, A. A. Salah and T. Gevers, Eds. Springer, 2011, pp. 227–253.
- [11] H. Kaya, D. Fedotov, D. Dresvyanskiy, M. Doyran, D. Mamontov, M. Markitantov, A. A. Akdag Salah, E. Kavcar, A. Karpov, and A. A. Salah, "Predicting depression and emotions in the crossroads of cultures, para-linguistics, and non-linguistics," in *Proc. 9th International on Audio/Visual Emotion Challenge and Workshop*, ser. AVEC '19, 2019, p. 27–35.
- [12] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.
- [13] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, Minnesota, USA,, 2007, pp. 1–8.
- [14] R. Remus, U. Quasthoff, and G. Heyer, "SentiWS - a publicly available German-language resource for sentiment analysis," in *LREC*, 2010.
- [15] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining," in *Lrec*, vol. 10, no. 2010, 2010, pp. 2200–2204.
- [16] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme Learning Machine for Regression and Multiclass Classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 2, pp. 513–529, 2012.
- [17] H. Wold, "Partial least squares," in *Encyclopedia of Statistical Sciences*, S. Kotz and N. L. Johnson, Eds. Wiley New York, 1985, pp. 581–591.
- [18] W. Zong, G.-B. Huang, and Y. Chen, "Weighted extreme learning machine for imbalance learning," *Neurocomputing*, vol. 101, pp. 229–242, 2013.
- [19] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.
- [20] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [22] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.
- [23] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," in *Proc. International Conference on Language Resources and Evaluation LREC-2018*, 2018.
- [24] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.
- [25] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. 8th International AAAI Conference on Weblogs and Social Media*, 2014.
- [26] S. Loria, "Textblob documentation," *Release 0.15*, vol. 2, 2018. [Online]. Available: <https://textblob.readthedocs.io/en/dev/>
- [27] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "Flair: An easy-to-use framework for state-of-the-art nlp," in *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 2019, pp. 54–59.
- [28] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF based framework for text categorization," *Procedia Engineering*, vol. 69, pp. 1356–1364, 2014.
- [29] D. Hiemstra, "A probabilistic justification for using tf × idf term weighting in information retrieval," *International Journal on Digital Libraries*, vol. 3, no. 2, pp. 131–139, 2000.
- [30] L.-P. Jing, H.-K. Huang, and H.-B. Shi, "Improved feature selection approach TFIDF in text mining," in *Proc. International Conference on Machine Learning and Cybernetics*, vol. 2. IEEE, 2002, pp. 944–946.
- [31] M. F. Porter *et al.*, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.
- [32] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. W. Schuller, "Snore sound classification using image-based deep spectrum features," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3512–3516.
- [33] S. Halfon, E. A. Oktay, and A. A. Salah, "Assessing affective dimensions of play in psychodynamic child psychotherapy via text analysis," in *International workshop on human behavior understanding*. Springer, 2016, pp. 15–34.