Complex & Intelligent Systems (2021) 7:1845–1854
https://doi.org/10.1007/s40747-020-00250-4

ORIGINAL ARTICLE

Cross corpus multi-lingual speech emotion recognition using ensemble learning

Wisha Zehra[1] · Abdul Rehman Javed[2] · Zunera Jalil[2] · Habib Ullah Khan[3] · Thippa Reddy Gadekallu[4]

Received: 26 September 2020 / Accepted: 3 December 2020 / Published online: 11 January 2021

Abstract

Receiving an accurate emotional response from robots has been a challenging task for researchers for the past few years. With the advancements in technology, robots like service robots interact with users of different cultural and lingual backgrounds. The traditional approach towards speech emotion recognition cannot be utilized to enable the robot and give an efficient and emotional response. The conventional approach towards speech emotion recognition uses the same corpus for both training and testing of classifiers to detect accurate emotions, but this approach cannot be generalized for multi-lingual environments, which is a requirement for robots used by people all across the globe. In this paper, a series of experiments are conducted to highlight an ensemble learning effect using a majority voting technique for cross-corpus, multi-lingual speech emotion recognition system. A comparison of the performance of an ensemble learning approach against traditional machine learning algorithms is performed. This study tests a classifier's performance trained on one corpus with data from another corpus to evaluate its efficiency for multi-lingual emotion detection. According to experimental analysis, different classifiers give the highest accuracy for different corpora. Using an ensemble learning approach gives the benefit of combining all classifiers' effect instead of choosing one classifier and compromising certain language corpus's accuracy. Experiments show an increased accuracy of 13% for Urdu corpus, 8% for German corpus, 11% for Italian corpus, and 5% for English corpus from with-in corpus testing. For cross-corpus experiments, an improvement of 2% when training on Urdu data and testing on German data and 15% when training on Urdu data and testing on Italian data is achieved. An increase of 7% in accuracy is obtained when testing on Urdu data and training on German data, 3% when testing on Urdu data and training on Italian data, and 5% when testing on Urdu data and training on English data. Experiments prove that the ensemble learning approach gives promising results against other state-of-the-art techniques.

Keywords Speech emotion recognition · Ensemble learning · Machine learning · Cross-corpus · Feature extraction · Cross-lingual

Introduction

Emotions help people communicate and understand others' opinions by conveying feelings and giving feedback to people [46]. Human speech renders a real and instinctive interface for communication with robots and is thus widely integrated into robots to interact with humans. Speechemotion recog-

✉ Thippa Reddy Gadekallu
thippareddy.g@vit.ac.in

Wisha Zehra
wishazehra1@gmail.com

Abdul Rehman Javed
abdulrehman.cs@au.edu.pk

Zunera Jalil
zunera.jalil@mail.au.edu.pk

Habib Ullah Khan
habibullah@qu.edu.qa

[1] National Center of Cyber Security, Air University, Islamabad, Pakistan

[2] Department of Cyber Security, Air University, Islamabad, Pakistan

[3] College of Business and Economics, Qatar University, Doha, Qatar

[4] School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India

nition is the act of attempting to understand the aspects of speech irrespective of the semantic contents and recognize the desired emotions using voice signals [19]. To enable robots to perceive a user's emotions accurately, a speech emotion recognition system can be integrated with simple speech recognition; however, the system should identify emotions for each individual independently of cultural and linguistic diversity.

Cross-corpus emotion recognition is the act of attempting to build classifiers that generalize across application scenarios and acoustic conditions and is highly relevant for constructing effective and practical speech emotion recognition systems [38]. Research has shown cross-corpus emotion recognition to be challenging for several reasons like differences in signal level, type of emotion elicitation, data scarcity, etc. Many researchers have tried to tackle these problems by creating their emotional corpus [20,27], trying out different feature sets [46], or using multiple machine learning models, but still, there is a lot of room for improvement. Ensemble learning helps to improve the performance of the machine learning models[17,29,33]. This prompts for further exploration of different techniques that can be used to improve cross-corpus speech emotion recognition that will enable the deployment of speech emotion recognition systems in real-life applications.

Human speech is so diverse and dynamic that no model can be reserved to be used forever [42]. This diversity of languages cause an imbalance of available datasets for emotion recognition for minority languages like Urdu or Sindhi vs. well-established majority languages like English. There is a need to establish a model that can be generalized for multi-lingual emotional data using the datasets available for us to use. The researchers need to examine how minority languages perform on models trained in majority languages.

Different machine learning algorithms [32] have been used to accurately classify emotions with-in the same corpus, but when applied to cross-corpus, the performance has been average. This highlights the fact that machine learning algorithms can detect emotions with-in the same corpus, but for cross-corpus, the researchers need to identify a way to utilize the ability of these machine learning algorithms to detect emotions to map out for cross-corpus data.

Existing studies [1,37,38] have either extracted an enormous amount of features that contribute to large computing times or have used a single machine learning algorithm [11,20], to classify emotions into its respective categories that have deprived us of using the information each classifier has to offer and instead rely on a single classifier which has proved to give lower accuracy than desired.

In this paper, researchers propose a speech emotion recognition system for robots that uses a combination of different audio features to detect accurate emotion with-in a corpus as well as cross-corpus using the ensemble learning approach. For this, the researchers use corpora in four different languages (Urdu, English, German, and Italian) and have chosen to conduct experiments with Urdu as the base language for various scenarios against the other three languages. The researchers investigate the effect of combining the classifiers used most popularly for speech emotion recognition by using a majority voting approach and demonstrate how it enhances cross-lingual emotion recognition.

In this paper, the researchers make the following contributions:

- Propose an effective ensemble learning approach to identify and detect cross-corpus emotions.
- Evaluate the effectiveness of the ensemble technique.
- Present a comparative analysis of conventional machine learning techniques: decision tree (J48), random forest (RF), and sequential minimal optimization (SMO) with an ensemble of these machine learning algorithms using majority voting.
- Ensemble learning approach effectively enhances the detection of emotion and achieves good accuracy on both with-in as well as cross-corpus data in comparison with conventional machine learning techniques.

The rest of the paper is organized as follows. "Related work" briefly covers the technical background and recent research on cross-corpus speech emotion recognition. "Proposed approach" presents an overview of our proposed approach of ensemble learning for cross-corpus speech emotion recognition. The experimental setup and results are articulated in "Evaluation and results". "Comparative analysis" presents comparative analysis and "Conclusion" concludes along with directions for future work.

Related work

Over the past 2 decades, there has been significant research on speaker-independent speech emotion recognition. This research has highlighted multiple factors that influence accurate detection of emotion; for example, the data set used, the features extracted, or the classifier used to predict emotions. Sailunaz et al. [36] described a detailed survey on multiple datasets available, the features extracted, and the models most used by multiple researchers. However, there is limited research available on multi-lingual cross-corpus speech emotion recognition. Initial studies exist on improving the sturdiness of multi-lingual speech emotion recognition by combining several emotional speech corpora within the training set and by that reducing the paucity of data [22].

The authors in [8] performed pilot experiments using support vector machines on four datasets of two different languages (German and English) to show the practicality of

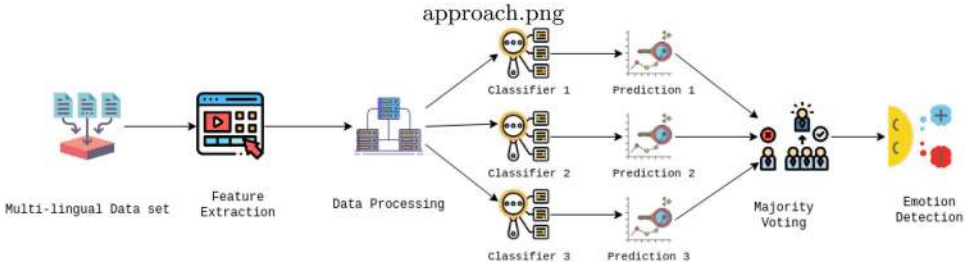

Fig. 1 Graphical representation of proposed ensemble learning approach for multi-lingual speech emotion recognition

cross-corpus emotion recognition. The authors in [37] have performed experiments using support vector machine on six datasets in three different languages (German, English, and Danish) and revealed the drawbacks of existing analysis and corpora. The authors in [1] developed an ensemble SVM for speech emotion recognition whose focus was on emotion recognition in never seen languages.

The authors in [35] identified a speaker's language to some extent and chose an appropriate model based on that knowledge. The authors in [44] chose an unsupervised learning approach to identify emotion on unlabeled data and found that unlabeled training data give approximately half of the gain that can be exacted from adding labeled training data. In [23], the authors used a three-layer model on corpora from three languages (German, Chinese, and Japanese) and found it accurate, yielding small errors. Li and Akagi [24] focused on choosing generalizable features from prosodic, spectral, and glottal waveform domains for multi-lingual speech emotion recognition. In [6], the authors used sparse autoencoders for feature transfer learning in speech emotion recognition. They used six standard databases and used the single-layer sparse autoencoder and trained this model on class-specific instances from the target domain, and then applied this representation to the source domain to reconstruct those data. This experimental approach improves the model's performance as compared to independent learning from every source domain. In [21], the authors used deep belief networks (DBN) for emotion recognition and found that networks with generalization power like deep belief networks are better than traditional discriminative networks like sparse auto en-coders, but this needs to be further investigated.

In [26], authors performed emotion recognition on two languages (English and French) and examined the performance of one model trained on multiple languages. Elbarougy et al. [7] examined the distinctions and commonalities of emotions in valence-activation space between three languages (Japanese, Chinese, and German) using 30 speakers and proved that emotions are almost similar between speakers speaking different languages. In [27], authors created a new emotional database named EmoSTAR in two languages (Turkish and English) and conducted cross-corpus tests with a German dataset using SVM. In [43], the authors performed experiments on three emotion corpora (Danish, Mandarin Chinese, and German) and achieved results that indicate universal cue in emotion expression regardless of language.

In [20], the authors created a new emotional database in Urdu language and performed experiments on three different language corpora (German, English, and Italian) using SVM classifier and evaluated the results of training and testing a model using different languages and found that adding some testing language data to the training data can improve performance. The authors in [45] used 1D and 2D CNN-LSTM networks to identify speech emotions. The authors in [40] analyzed the effect noise removal techniques have on SER systems. The authors in [11] performed transfer learning and multi-task learning experiments and found that traditional machine learning models may function as well as deep learning models [2,41] for speech emotion recognition given the researchers choose the right input feature.

Proposed approach

Many factors influence the accurate detection of emotion in a cross-corpus setting. The dataset used, the features extracted from the audio signals, and the classifiers used to detect emotion all factors can significantly influence your results. Figure 1 summarizes our approach for multi-lingual speech emotion recognition. This study works on four corpora (SAVEE, URDU, EMO-DB, and EMOVO) that give a diversity of languages (English, Urdu, German, and Italian) to test for multi-lingual speech emotion recognition. To ensure that researchers have the same class labels for every dataset, this study uses the binary valence (positive and negative) approach, as presented in Table 1. The proposed approach works by extracting a combination of spectral and prosodic features from raw audio files to feed into the classifier. The Ensemble learning approach through majority voting is used to train the model to classify emotions into their respective category accurately. Further details on the selected databases, speech features extracted, and the Ensemble classifiers are presented below.

Table 1 Corpora information

References	Corpus	Lang	Spk	Utt	Cat	Positive valence	Negative valence
[13]	SAVEE	English	4	480	Acted	Neutral, happiness, surprise	Anger, sadness, fear, disgust
[20]	Urdu	Urdu	38	400	Acted	Neutral, happiness	Anger, sadness
[3]	EMO-DB	German	10	497	Acted	Neutral, happiness	Anger, sadness, fear, boredom, disgust
[5]	EMOVO	Italian	6	588	Natural	Neutral, happiness, surprise	Anger, sadness, fear, disgust

Utt Utterences, *Spk* speakers, *Lang* language, *Cat* category

Speech emotion databases

For multi-lingual speech emotion recognition, the data should be diverse. For this study, four datasets, each with a different language, are selected based on their recording environments, the categories of emotion classes available, and the balance between positive and negative valence classes.

SAVEE

The surrey audio–visual expressed emotion (SAVEE) database [13] was recorded from four male English speakers. Emotion is categorized into seven discrete categories: anger, disgust, happy, sad, fear, neutral, and surprise. There are a total of 120 utterances for each speaker. The audio has been recorded in a controlled environment and is acted out by the speakers. The corpus is publicly available[1] for research.

Urdu

The Urdu database [20] contains audio recordings collected from Urdu TV talk shows, consisting of 400 recordings from 38 speakers (27 male, 11 female). The data are collected for four basic emotions: anger, happy, sad, and neutral. This corpus contains natural emotional excerpts from real and unscripted discussions between different guests of TV talk shows. The dataset is publicly available [2] for research.

EMO-DB

The Berlin database of emotional speech [3] is a German database containing speech audios from 10 actors (5 male, 5 female). The data consist of 10 German sentences recorded in anger, boredom, disgust, fear, happiness, sadness, and neutral. This database has 497 annotated utterances and has been recorded in a studio with trained actors to get an appropriate emotional response. This corpus is available[3] for research purposes.

EMOVO

EMOVO is an Italian speech emotion database [5] that consists of recordings from 6 actors (3 male, 3 female) simulating 7 emotional states: disgust, fear, anger, joy, surprise, sadness, and neutral. There are 14 sentences uttered for each emotion and have a total of 588 annotated audio recordings. These audio recordings were recorded in a studio by trained actors and are the first emotional database for the Italian language, and are available online.[4]

Feature extraction

The authors in [11] deduced that choosing the right input features can be the key to efficient recognition of emotion [30]. This work experimented with different types of features, both spectral as well as prosodic, against each dataset. Mel-frequency Cepstral Coefficients (MFCC) are among the most widely used features for speech and emotion recognition. To generate MFCCs, researchers use Librosa [25] Python library. This study considers the first 20 sets of MFCCs for experimentation. Aside from MFCCs, Spectral (Roll-off, flux, centroid, bandwidth), Energy (Root-mean-square energy), Raw Signal (Zero crossing rate), Pitch (Fundamental frequency), and Chroma features are also used for experimentation. Each feature is calculated at every 0.02 s of the audio files. Then, the researchers use the most common statistical approach and take the median of all the values calculated at each frame to constitute the value for the corresponding feature. Table 2 describes the features extracted from each feature group. A total of 28 features are extracted against each audio file, and the results are stored in a CSV file.

To test the performance of the selected functions as input functions, this work uses a different set of features, i.e., eGeMAPS, which consists of 88 features connected to energy, spectrum, frequency, Cepstral, and dynamic information. Details on these features can be found in [10]. To extract eGeMAPS features, the researchers use openSMILE toolkit [9] and save results in a CSV file.

[1] http://kahlan.eps.surrey.ac.uk/savee/Download.html.

[2] https://github.com/siddiquelatif/URDU-Dataset.

[3] http://www.emodb.bilderbar.info/download/.

[4] https://mega.nz/file/b5tSDDAK#-saGyczbcMWl-jXg4RHon7xU_pc8QHg0sQtikmIg2c4.

Table 2 Features extracted

Feature group	Features in group
Cepstral	MFCC 0 - 19
Spectral	Flux, roll-off point, centroid, bandwidth
Raw signal	Zero crossing rate
Pitch	Fundamental frequency F0, Chroma
Signal energy	Root mean square

Preprocessing

An imbalanced dataset causes machine learning algorithms to under-perform [14,18,28,31]. The synthetic minority oversampling technique (SMOTE) [4,15,16] is a powerful approach to tackle the class imbalance problem. After feature extraction [34], SMOTE is used to balance the instances in each class for our experimentation. After feature extraction, the data have a wide range of values that need to be converted to a common scale for our classifiers to perform well. Data normalization is performed to scale the values of our features between 0 and 1 [12].

Classification models and parameter setting

For experimentation, this approach uses Support Vector Machines (SVM) to provide good classification results even if the researchers have a small dataset. SVM is known to perform well in higher dimension data, which can usually be the case when working with audio data, and it has been widely used for speech emotion recognition . The proposed approach uses SVM with puk Kernel, complexity 1.0, and pairwise multi-class discrimination based on Sequential Minimal Optimization. Furthermore, this study uses Random Forest, another benchmark classifier used widely for classification problems. This work uses the Random Forest with 10 trees for experimentation. Decision Tree (J48) is also used to classify data into its respective category. Decision Trees are used with a confidence factor of 0.25 for pruning, and the minimum number of instances per leaf was set to 2. Finally, this study uses an ensemble learning approach through majority voting. The proposed approach utilizes SMO, RF, and J48 classifier in an ensemble for cross-corpus emotion recognition.

Evaluation and results

This study conducts multiple experiments by setting Urdu as the base language to test against the remaining three languages (English, German, and Italian). The researchers use the 'leave-one-speaker-out' scheme to split our data into training and testing sets. The researchers use accuracy, precision, recall, and f-score to evaluate the proposed ensemble model's performance.

Figure 2 gives an overview of the results achieved. This work experiment uses multiple machine learning languages and an ensemble learning approach, described below.

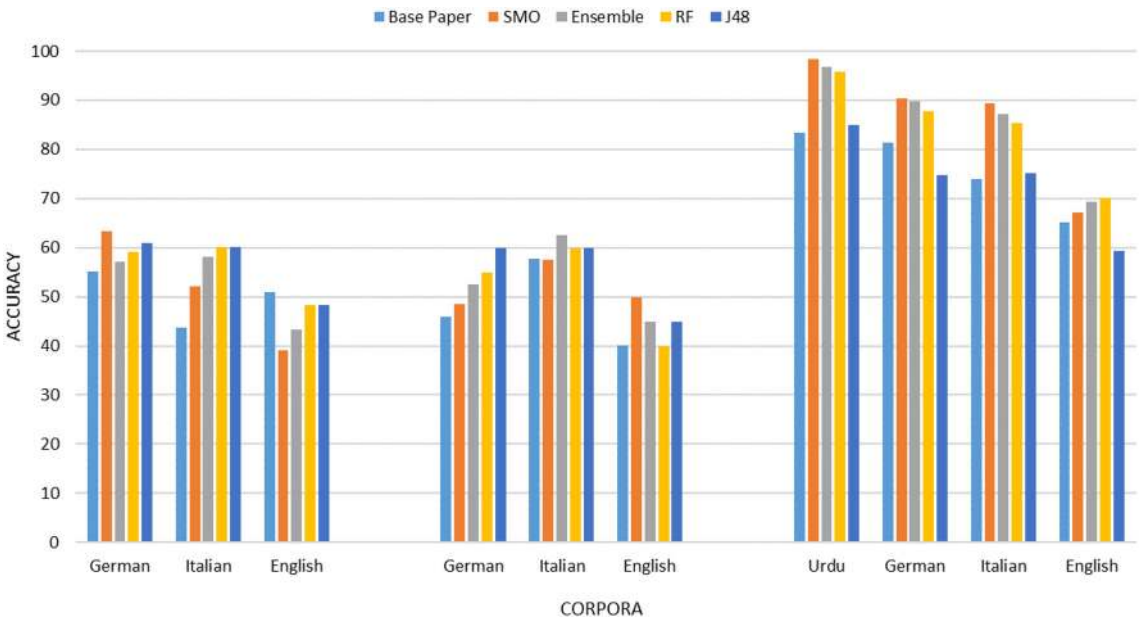

Fig. 2 Results achieved using Urdu as training set, Urdu as testing set, and with-in corpus experiments

With-in corpus experiments

This work conducts with-in corpus experiments to establish a baseline for features with the researchers' corpora using classifiers' set. For this experiment, the researchers use training and testing data from the same corpus. This helps to understand how well the models can perform given a certain corpus. As depicted in Fig. 3, the Urdu corpus gives impressive results as SMO gave an accuracy of 98.5% followed by the ensemble with an accuracy of 96.75%. For the EMO-DB (German) corpus, SMO gave an accuracy of 90.4% followed closely by the ensemble learning approach which gives an accuracy of 89.75%. For SAVEE (English), corpus RF gives the highest accuracy of 70.14%, while ensemble learning gives 69.31%. Finally, for EMOVO (Italian) database, SMO gives an accuracy of 89.41% followed by an ensemble learning approach with an accuracy of 87.14%. From this experiment, the researchers observe that no matter which algorithm gives the highest accuracy, ensemble learning stood second and not by much margin. SMO may perform better for some corpus, while RF may be best for another. The researchers cannot generalize one classifier working best for cross-corpus data. On the other hand, the ensemble learning approach gives us comparable results that can be used for cross-corpus speech emotion recognition without having to compromise on a lower accuracy rate for some language.

Cross-corpus experiments

For this set of experiments, the experiment pattern of [20] is followed. This work first uses Urdu data for training the model and testing it against the three western languages (English, German, and Italian). This study performs experiments using the three machine learning algorithms (SMO, RF, and J48) and the ensemble learning approach. Tables 3, 4, and 5 depicts the performance of the classifier against each corpus. It was interesting to note that a different classifier was seen to perform the best for each corpus. When testing with data from EMO-DB (German) corpus, the classifier SMO with puk kernel performs the best and gives an accuracy of 63% while the other classifiers give a lower accuracy. When testing with data from EMOVO (Italian) corpus, random forest (RF) performs the best and gives an accuracy of 60.02%, while the other classifiers give a lower accuracy. Finally, when testing on SAVEE (English) corpus, J48 gives an accuracy of 48.34%, which is again higher than the other classifiers. This observation leads to a question: which classifier should be used to implement a multi-lingual speech emotion recognition system? The ensemble learning approach may not give the best accuracy, but it shows promising results when trained using Urdu data and tested against the other three corpora. It answers the question of which classifier to use by combining the effect of all three classifiers. This ensemble uses a majority voting approach that ensures accuracy for a cross-corpus model.

For the next set of experiments, the proposed approach uses EMO-DB corpus for training and Urdu data for testing. This study evaluates all classifiers and gets an accuracy of 60% from the J48 classifier, while other classifiers give moderate accuracy, as shown in Table 6. This study then uses EMOVO (Italian) corpus for training the models and testing them against Urdu data. In this case, the Ensemble gives the highest accuracy of 62.5%, while individual classifiers give lower accuracy, as shown in Table 7. Finally, this study trains the models using SAVEE (English) corpus while tests it using Urdu data. SMO classifier gives the highest accuracy of 50%, while the other classifiers give an inferior performance, as shown in Table 8.

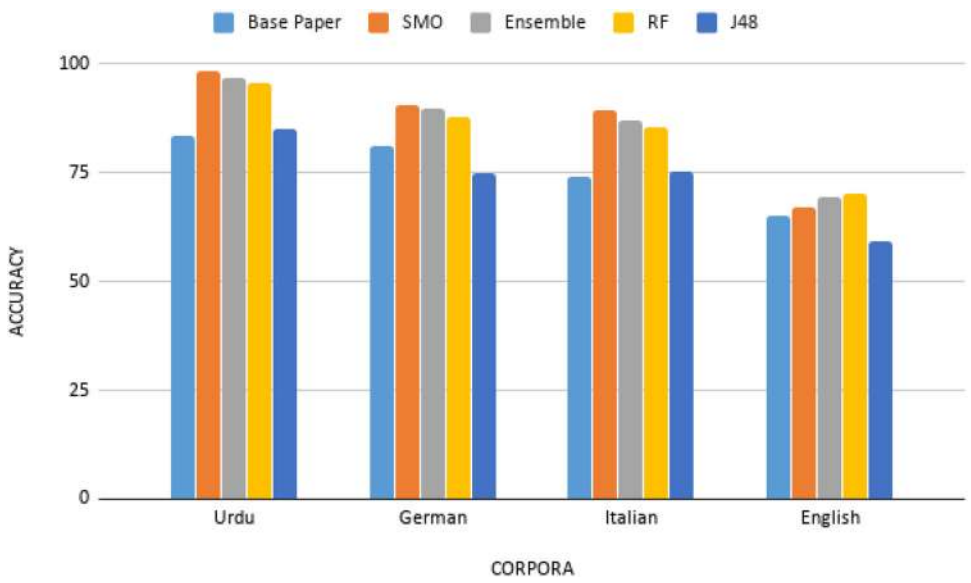

Fig. 3 With-in-corpus results

Table 3 Training on Urdu corpus, testing on Italian corpus

Classifier	Precision	Recall	F-score	Accuracy
J48	0.59	0.60	0.59	60.20%
SMO	0.46	0.52	0.45	52.04%
RF	0.59	0.60	0.55	60.20%
Ensemble	0.38	0.58	0.53	58.16%

Table 4 Training on Urdu corpus, testing on German corpus

Classifier	Precision	Recall	F-score	Accuracy
J48	0.60	0.61	0.61	61.22%
SMO	0.60	0.63	0.58	63.26%
RF	0.56	0.59	0.56	59.18%
Ensemble	0.54	0.57	0.55	57.14%

Table 5 Training on Urdu corpus, testing on English corpus

Classifier	Precision	Recall	F-score	Accuracy
J48	0.45	0.48	0.38	48.34%
SMO	0.34	0.39	0.34	39.16%
RF	0.47	0.48	0.44	48.34%
Ensemble	0.38	0.43	0.36	43.34%

Table 6 Training on German corpus, testing on Urdu corpus

Classifier	Precision	Recall	F-score	Accuracy
J48	0.60	0.60	0.60	60%
SMO	0.46	0.49	0.37	48.5%
RF	0.63	0.55	0.46	55%
Ensemble	0.75	0.52	0.38	52.5%

Table 7 Training on Italian corpus, testing on Urdu corpus

Classifier	Precision	Recall	F-score	Accuracy
J48	0.60	0.60	0.59	60%
SMO	0.67	0.57	0.50	57.5%
RF	0.63	0.60	0.57	60%
Ensemble	0.67	0.62	0.59	62.5%

Table 8 Training on English corpus, testing on Urdu corpus

Classifier	Precision	Recall	F-score	Accuracy
J48	0.45	0.45	0.45	45%
SMO	0.50	0.50	0.46	50%
RF	0.39	0.40	0.39	40%
Ensemble	0.44	0.45	0.43	45%

This set of experiments also support the observation that no one classifier was performing the best for every scenario.

Comparative analysis

To analyze the efficacy of the proposed approach, this study compares the results with a distinguished research conducted by [20] whose pattern of experimentation was followed in this study. The authors have extracted eGeMAPS [10] features from their raw audio data. They have used SVM with a gaussian kernel for classifying data into their respective categories. Figure 4 compares the accuracy of the proposed ensemble learning approach with the referred

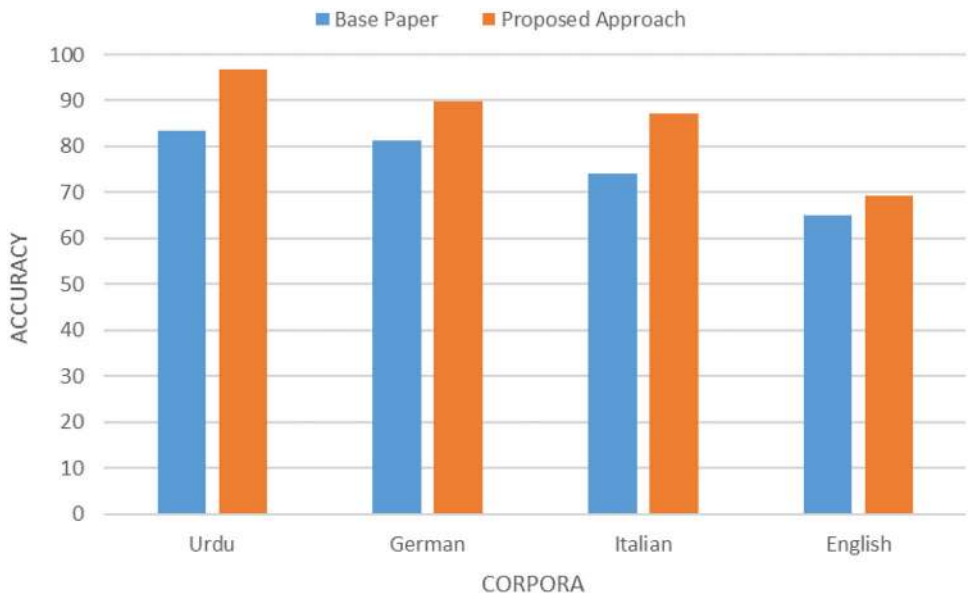

Fig. 4 Performance comparison of proposed approach with referred paper

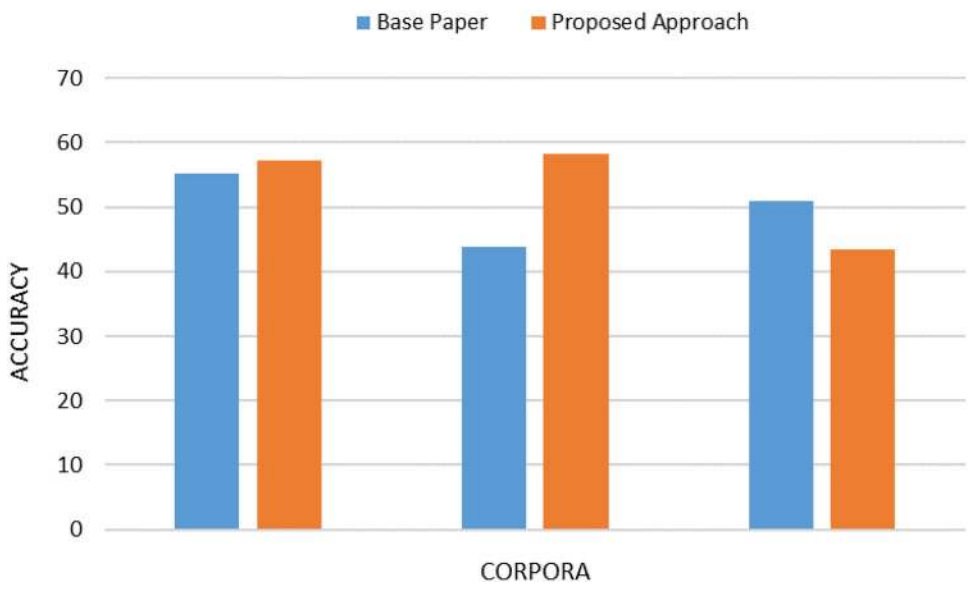

Fig. 5 Performance comparison of proposed approach with referred paper setting Urdu data as training data and testing on data from other languages

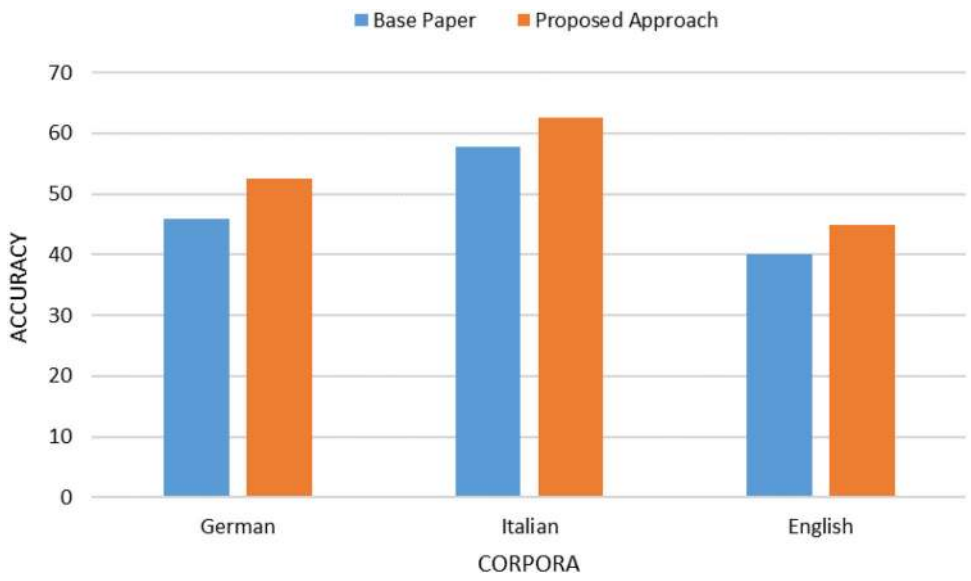

Fig. 6 Performance comparison of the proposed approach with the referred paper setting Urdu data as testing data and training on data from other languages

research paper's accuracy. For the Urdu database, the ensemble learning approach shows increased accuracy by 13%. For EMO-DB, the accuracy increased by 8% using ensemble learning. For EMOVO (Italian) corpus, the ensemble learning improved the accuracy by 11%. Finally, for SAVEE (English) corpus, almost a 5% increase in accuracy was achieved using the ensemble learning approach.

Figures 5 and 6 present an overview of cross-corpus comparison. When training against Urdu corpus, EMO-DB (German) and EMOVO (Italian) give us an increased accuracy of 2% and 15%, respectively. For the SAVEE corpus, this study observes a decline of 6% using the ensemble learning approach. When testing using the Urdu corpus, this work achieves an increased accuracy of 7%, 3%, and 5% for German, Italian, and English corpus, respectively.

Conclusion

The paradigm shift from textual to more intuitive control mechanisms like speech in human–robot interaction (HRI) has opened several research areas, including speech emotion recognition. A lot of past research for speech emotion recognition has been focused on using the data from the same corpus for both training and testing. This study proposed an ensemble learning technique through majority voting to tackle emotions in multiple languages and enable the robots to perform globally. It is observed that different classifiers worked differently for different languages, which raised the question of which classifier works best for all languages. The Ensemble learning approach, which uses the three most popular machine learning algorithms and implements a majority voting scheme, gave comparable results for all languages.

This finding can be very helpful for developing an emotion recognition system for robots designed to handle customers from all corners of the globe [39]. It will enable the robots to interact with customers smartly with emotional intelligence, which can have a huge impact on the way the world interacts with robots. The researchers plan to explore more machine learning algorithms to be used in an ensemble in the future. To enable the application of our research in real-life scenarios, the researchers want to experiment with different speech databases containing audios recorded in a natural environment. Moreover, the researchers plan to analyze the effect of using different ensemble techniques and achieving higher accuracy rates. The most challenging task for future researchers would be finding corpora for different languages in the natural environment as there are not many readily available. Second, selecting algorithms that perform consistently for all languages in both natural and recorded environments.

Compliance with ethical standards

Conflict of interest The authors declare that they do not have any conflicts of interests.

References

1. Albornoz EM, Milone DH (2015) Emotion recognition in never-seen languages using a novel ensemble method with emotion profiles. IEEE Trans Affect Comput 8(1):43–53
2. Bhattacharya S, Maddikunta PKR, Pham QV, Gadekallu TR, Chowdhary CL, Alazab M, Piran MJ, et al. (2020) Deep learning and medical image processing for coronavirus (covid-19) pandemic: a survey. Sustain Cities Soc 102589. https://doi.org/10.1016/j.scs.2020.102589
3. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B (2005) A database of German emotional speech. In: Proceeding of the INTERSPEECH, Lisbon, Portugal, pp 1517–1520
4. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
5. Costantini G, Iaderola I, Paoloni A, Todisco M (2014) EMOVO corpus: an Italian emotional speech database. In: Proceedings of the ninth international conference on language resources and evaluation (LREC'14), European language resources association (ELRA), Reykjavik, Iceland, pp 3501–3504. http://www.lrec-conf.org/proceedings/lrec2014/pdf/591_Paper.pdf. Accessed 1 Oct 2020
6. Deng J, Zhang Z, Marchi E, Schuller B (2013) Sparse autoencoder-based feature transfer learning for speech emotion recognition. In: 2013 humaine association conference on affective computing and intelligent interaction. IEEE, pp 511–516 ACII 2013 6681481
7. Elbarougy R, Xiao H, Akagi M, Li J (2014) Toward relaying an affective speech-to-speech translator: cross-language perception of emotional state represented by emotion dimensions. Oriental COCOSDA 2014-17th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment / CASLRE (Conference on Asian Spoken Language Research and Evaluation) 7051419
8. Eyben F, Batliner A, Schuller B, Seppi D, Steidl S (2010) Cross-corpus classification of realistic emotions–some pilot experiments. In: Proceedings of 7th international conference on language resources and evaluation (LREC 2010), Valletta, Malta
9. Eyben F, Wöllmer M, Schuller B (2010) OpenSMILE - the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM International Conference on Multimedia, MM 2010, (Florence, Italy), pp 1459–1462
10. Eyben F, Scherer KR, Schuller BW, Sundberg J, André E, Busso C, Devillers LY, Epps J, Laukka P, Narayanan SS et al (2015) The Geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE Trans Affect Comput 7(2):190–202
11. Goel S, Beigi H (2020) Cross lingual cross corpus speech emotion recognition. arXiv preprint arXiv:2003.07996
12. Imtiaz SI, ur Rehman S, Javed AR, Jalil Z, Liu X, Alnumay WS (2020) Deepamd: detection and identification of android malware using high-efficient deep artificial neural network. Future Gener Comput Syst 115:844–856
13. Jackson P, Haq S (2014) Surrey audio-visual expressed emotion (savee) database. University of Surrey, Guildford
14. Javed AR, Beg MO, Asim M, Baker T, Al-Bayatti AH (2020) Alphalogger: detecting motion-based side-channel attack using smartphone keystrokes. J Ambient Intell Humaniz Comput 1–14. https://doi.org/10.1007/s12652-020-01770-0
15. Javed AR, Fahad LG, Farhan AA, Abbas S, Srivastava G, Parizi RM, Khan MS (2020) Automated cognitive health assessment in smart homes using machine learning. Sustain Cities Soc. https://doi.org/10.1007/s12652-020-01770-0
16. Javed AR, Sarwar MU, Khan S, Iwendi C, Mittal M, Kumar N (2020) Analyzing the effectiveness and contribution of each axis of tri-axial accelerometer sensor for accurate activity recognition. Sensors 20(8):2216
17. Javed AR, Usman M, Rehman SU, Khan MU, Haghighi MS (2020) Anomaly detection in automated vehicles using multistage attention-based convolutional neural network. IEEE Trans Intell Transport Syst. https://doi.org/10.1109/TITS.2020.3025875
18. Kaur D, Aujla GS, Kumar N, Zomaya AY, Perera C, Ranjan R (2018) Tensor-based big data management scheme for dimensionality reduction problem in smart grid systems: Sdn perspective. IEEE Trans Knowl Data Eng 30(10):1985–1998
19. Khan MU, Javed AR, Ihsan M, Tariq U (2020) A novel category detection of social media reviews in the restaurant industry. Multimed Syst. https://doi.org/10.1007/s00530-020-00704-2
20. Latif S, Qayyum A, Usman M, Qadir J (2018) Cross lingual speech emotion recognition: Urdu vs. western languages. In: Proceedings - 2018 International Conference on Frontiers of Information Technology, FIT 2018 8616972, pp 88–93
21. Latif S, Rana R, Younis S, Qadir J, Epps J (2018) Cross corpus speech emotion classification: an effective transfer learning technique. arXiv preprint arXiv:1801.06353

22. Lefter I, Rothkrantz LJ,Wiggers P, Van Leeuwen DA (2010) Emotion recognition from speech by combining databases and fusion of classifiers. In: 13th International Conference on Text, Speech and Dialogue, Czech Republic, Vol 6231, pp 353–360
23. Li X, Akagi M (2016) Multilingual speech emotion recognition system based on a three-layer model. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp 3608–3612
24. Li X, Akagi M (2018) A three-layer emotion perception model for valence and arousal-based detection from multilingual speech. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, pp 3643–3647
25. McFee B, Raffel C, Liang D, Ellis DP, McVicar M, Battenberg E, Nieto O (2015) librosa: audio and music signal analysis in python. In: Proceedings of the 14th python in science conference, vol 8
26. Neumann M et al (2018) Cross-lingual and multilingual speech emotion recognition on English and French. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 5769–5773
27. Parlak C, Diri B, Gürgen F (2014) A cross-corpus experiment in speech emotion recognition. In: SLAM@ INTERSPEECH, pp 58–61
28. Patel H, Singh Rajput D, Thippa Reddy G, Iwendi C, Kashif Bashir A, Jo O (2020) A review on classification of imbalanced data for wireless sensor networks. Int J Distrib Sens Netw 16(4):1550147720916404
29. Reddy GT, Bhattacharya S, Ramakrishnan SS, Chowdhary CL, Hakak S, Kaluri R, Reddy MPK (2020) An ensemble based machine learning model for diabetic retinopathy classification. In: International conference on emerging trends in information technology and engineering, ic-ETITE 2020 9077904. IEEE, pp 1–6
30. Reddy GT, Reddy MPK, Lakshmanna K, Kaluri R, Rajput DS, Srivastava G, Baker T (2020) Analysis of dimensionality reduction techniques on big data. IEEE Access 8:54776–54788
31. Reddy T, Bhattacharya S, Maddikunta PKR, Hakak S, Khan WZ, Bashir AK, Jolfaei A, Tariq U (2020) Antlion re-sampling based deep neural network model for classification of imbalanced multimodal stroke dataset. Multimed Tools Appl. https://doi.org/10.1007/s11042-020-09988-y
32. Rehman ZU, Zia MS, Bojja GR, Yaqub M, Jinchao F, Arshid K (2020) Texture based localization of a brain tumor from MR-images by using a machine learning approach. Med Hypotheses. https://doi.org/10.1016/j.mehy.2020.109705
33. Rehman JA, Jalil Z, Atif MS, Abbas S, Liu X (2020) Ensemble adaboost classifier for accurate and fast detection of botnet attacks in connected vehicles. Trans Emerg Telecommun Technol. https://doi.org/10.1002/ett.4088
34. RM SP, Maddikunta PKR, Parimala M, Koppu S, Reddy T, Chowdhary CL, Alazab M (2020) An effective feature engineering for dnn using hybrid pca-gwo for intrusion detection in iomt architecture. Comput Commun 8:54776–54788
35. SaghaH, Matejka P, Gavryukova M, Povolnỳ F,Marchi E, Schuller BW (2016) Enhancing multilingual recognition of emotion in speech by language identification. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, pp 2949–2953
36. Sailunaz K, Dhaliwal M, Rokne J, Alhajj R (2018) Emotion detection from text and speech: a survey. Soc Netw Anal Min 8(1):28
37. Schuller B, Vlasenko B, Eyben F, Wöllmer M, Stuhlsatz A, Wendemuth A, Rigoll G (2010) Cross-corpus acoustic emotion recognition: variances and strategies. IEEE Trans Affect Comput 1(2):119–131
38. Schuller B, Zhang Z, Weninger F, Rigoll G (2011) Using multiple databases for training in emotion recognition: to unite or to vote? In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, pp 1553–1556
39. Shrivastava R, Kumar P, Tripathi S, Tiwari V, Rajput DS, Gadekallu TR, Suthar B, Singh S, Ra IH (2020) A novel grid and place neuron's computational modeling to learn spatial semantics of an environment. Appl Sci 10(15):5147
40. Triantafyllopoulos A, Keren G, Wagner J, Steiner I, Schuller BW (2019) Towards robust speech emotion recognition using deep residual networks for speech enhancement. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, pp 1691–1695
41. Venkatraman S, Alazab M, Vinayakumar R (2019) A hybrid deep learning image-based analysis for effective malware detection. J Inf Secur Appl 47:377–389
42. Wang D, Zheng TF (2015) Transfer learning for speech and language processing. In: Asia-Pacific signal and information processing association annual summit and conference, APSIPA ASC 2015 7415532, pp 1225–1237
43. Xiao Z,Wu D, Zhang X, Tao Z (2016) Speech emotion recognition cross language families:Mandarin vs. western languages. In: PIC 2016 - Proceedings of the 2016 IEEE international conference on progress in informatics and computing 7949505, pp 253–257
44. Zhang Z, Weninger F, Wöllmer M, Schuller B (2011) Unsupervised learning in cross-corpus acoustic emotion recognition. In: IEEE workshop on automatic speech recognition and understanding, ASRU 2011, Proceedings 6163986, pp 523–528
45. Zhao J, Mao X, Chen L (2019) Speech emotion recognition using deep 1d & 2d cnn lstm networks. Biomed Signal Process Control 47:312–323
46. Zvarevashe K, Olugbara O (2020) Ensemble learning of hybrid acoustic features for speech emotion recognition. Algorithms 13(3):70