Project proposal:

Classify multilingual emotional speech audio by valence

Rubric

Problem statement

Emotion recognition is an important part of natural language understanding. Conversational agents accepting voice input have already been deployed in many contexts such as healthcare [1] or customer service [2] where empathic responses improve the quality of services provided. A cross-lingually or multilingually trained classifier can be especially useful when little training data is available for a particular target language [3].

This is a supervised classification problem to predict the valence of human speech using acoustic features of speech samples.

Previous work

[4] combined five corpora (German, Italian, and English variants) and attained F1 scores between 89% and 98% and corresponding accuracy scores between 92% and 98% for valence classification. Using a model trained on English and French data, [3] achieved an unweighted average recall of 61.73% (English) and 49.33% (French) for valence. [5] trained a model on English, German, Italian, and Urdu to obtain an unweighted average recall score of 70.98% for binary (negative and non-negative) valence classification in Urdu. Others have developed multilingual speech emotion recognition systems to classify utterances into affective categories (e.g., happy, sad, angry, neutral, etc.) [6]–[9].

Contributions of this project

This project builds on prior research by incorporating a broader set of multilingual data: 11 English datasets, 9 datasets in non-English languages, and 3 datasets providing both English and non-English speech samples for a total of 24 data sources in nine languages.

In addition, a deep learning approach will be adopted to compare results to the stacked ensemble (random decision forest, AdaBoost, logistic regression, and gradient boosting machine) used in [4].

Computational resources

My local machine provides 16 gigabytes of random access memory (RAM) and an Intel i7 processor with 4 cores and base speed of 2.6 gigahertz as well as a graphics processing unit providing 4 gigabytes of additional memory. Access to additional computational resources (e.g., Amazon Web Services) is provided by Springboard and may be utilized as necessary.

Project deliverables

Data, process (including cleaning, feature engineering, and training), and results will be documented and summarized in a Medium post. In addition, a web page will be developed where interested users may upload their own audio clips to see the model's predicted classification as well as to collect out-of-sample speech data that may be used to further develop the model.

Datasets

English audio samples with emotion labels may be sourced from the Carnegie Mellon University Let's Go Spoken Dialogue Corpus [10]–[11], Crowd-sourced Emotional Multimodal Actors Dataset [12]–[13], the Electromagnetic Articulography Database [14], the EmoReact dataset [15], the eNTERFACE '05 Audio-Visual Emotion Database [16], the JL Corpus [17], the Multimodal EmotionLines Dataset [18]–[19], the Ryerson Audio-Visual Database of Emotional Speech and Song [20], the Surrey Audio-Visual Expressed Emotion Database [21], the Toronto Emotional Speech Set [22], and the Variably Intense Vocalizations of Affect and Emotion Corpus [23].

Similar spoken corpora with emotion labels may be obtained for Arabic (Arabic Natural Audio Dataset) [25], Estonian (Estonian Emotional Speech Corpus) [26], French (French Emotional Speech Database - Oréau) [27] and Canadian (Québec) French (Canadian French Emotional Speech Database) [28], German (Berlin Database of Emotional Speech) [29], Greek (Acted Emotional Speech Dynamic Database) [30]–[31], Persian (Sharif Emotional Speech Database) [32], Turkish (BAUM-1) [33], and Urdu (Urdu Language Speech Dataset) [5].

Three datasets containing non-English samples in addition to English samples are available: BAUM-2 (Turkish) [34], the Emotional Speech Dataset (Mandarin Chinese) [35], and the Emotional Voices Database (Belgian French) [36].

The end-user license agreements of BAUM-1 [33], BAUM-2 [34], the EmoReact dataset [15], and the Surrey Audio-Visual Expressed Emotion Database [21] do not allow for distribution in any way.

Each dataset was created with different methods, but they share common features that made them suitable for this project:

1. Audio data (or video data with audio) of natural human speech at the word or utterance level from a single speaker. These were variously obtained via spontaneous participant elicitation (e.g., the eNTERFACE '05 Audio-Visual Emotion Database [16] or Estonian Emotional Speech Corpus [26]), acted speech (e.g., Ryerson Audio-Visual Database of Emotional Speech and Song [20] or Acted Emotional Speech Dynamic Database [30]–[31]), or television media samples (e.g., the Urdu Language Speech Dataset [5] or Multimodal EmotionLines Dataset [18]–[19]).

2. A single unambiguous valence classification of either positive, negative, or neutral per audio sample — explicitly labeled or directly inferable from the conventional valences of basic emotion categories. For instance, samples colored with anger, disgust, fear, or sadness would all be considered negatively valenced. Joy is considered positively valenced. Because the valence of surprise is ambiguous [37], it would not be considered a valid label for this project. Datasets dealing with sentiment but not emotions would be omitted from consideration.

3. Demonstrated academic or practical application in some context (e.g., conference or journal publication, Kaggle, etc.).

4. Public or free-use non-commercial access.

Information about speaker gender is available for the above datasets as well (although not always explicitly encoded). All speakers in all datasets were adults with the exception of the EmoReact dataset [15], which featured children's English, and the Canadian French Emotional Speech Database [28] in which one of the actors was under 18 years old at time of recording.

Many other datasets such as some of those listed in [38] did not meet all the above criteria, required an active academic affiliation, were paywalled, or were otherwise inaccessible.

References

[1] L. Laranjo, A. G. Dunn, H. Y. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. S. Lau, and E. Coiera, "Conversational agents in healthcare: A systematic review," J. Amer. Med. Inform. Assoc., vol. 25, no. 9, pp. 1248–1258, Jul. 11, 2018, doi: https://doi.org/10.1093/jamia/ocy072.

[2] U. Gnewuch, S. Morana, and A. Maedche, "Towards designing cooperative and social conversational agents for customer service," in Proc. 38th Int. Conf. Inf. Syst., Seoul, South Korea, Dec. 10–13, 2017. Accessed: Mar. 3, 2021. [Online]. Available: https://chatbotresearch.com/wp-content/uploads/2018/06/icis2017.pdf

[3] M. Neumann and N. T. Vu, "Cross-lingual and multilingual speech emotion recognition on English and French," in Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process., Calgary, AB, Canada, Apr. 15–20, 2018, pp. 5769–5773. doi: https://doi.org/10.1109/ICASSP.2018.8462162.

[4] K. Zvarevashe and O. O. Olugbara, "Recognition of cross-language acoustic emotional valence using stacked ensemble learning," Algorithms, vol. 13, no. 10, p. 246, Sep. 27, 2020, doi: https://doi.org/10.3390/a13100246.

[5] S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Cross lingual speech emotion recognition: Urdu vs. Western languages," 2020, arXiv:1812.10411. Accessed Feb. 10, 2021. [Online]. Available: https://arxiv.org/pdf/1812.10411.pdf

[6] R. Elbarougy and M. Akagi, "Cross-lingual speech emotion recognition system based on a three-layer model for human perception," in 2013 Asia-Pacific Signal and Inf. Process. Assoc. Annu. Summit and Conf., Kaohsiung, Taiwan, Oct. 29–Nov. 1, 2013, pp. 1-10. doi: https://doi.org/10.1109/APSIPA.2013.6694137.

[7] P. Heracleous and A. Yoneyama, "A comprehensive study on bilingual and multilingual speech emotion recognition using a two-pass classification scheme," PLoS ONE, vol. 14, no. 8, p. e0220386, Aug. 15, 2019, doi: https://doi.org/10.1371/journal.pone.0220386.

[8] X. Li and M. Akagi, "Multilingual speech emotion recognition system based on a three-layer model," in Proc. INTERSPEECH 2016, San Francisco, CA, USA, Sep. 8–12, 2016, pp. 3608–3612. doi: https://doi.org/10.21437/Interspeech.2016-645.

[9] X. Li and M. Akagi, "Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model," Speech Communication, vol. 110, pp. 1–12, Jul. 2019, doi: https://doi.org/10.1016/j.specom.2019.04.004.

[10] A. Schmitt, S. Ultes, and W. Minker, "A parameterized and annotated spoken dialog corpus of the CMU Let's Go bus information system," in Int. Conf. Lang. Resour. and Eval., Istanbul, Turkey, May 2012, pp. 3369–3373. Accessed: Feb. 8, 2021. Available: https://www.academia.edu/21586940/A_Parameterized_and_Annotated_Spoken_Dialog_Corpus_of_the_CMU_Lets_Go_Bus_Information_System

[11] S. Ultes, A. Schmitt, M. J. P. Sánchez, and W. Minker, "Analysis of an extended interaction quality corpus," in Natural Lang. Dialog Syst. and Intell. Assistants, G. G. Lee, H. K. Kim, M. Jeong, and J.-H. Kim, Eds., Cham, Switzerland: Springer Int. Publishing, 2015, pp. 41–52. doi: https://doi.org/10.1007/978-3-319-19291-8_4.

[12] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset," IEEE Trans. Affect. Comput., vol. 5, no. 4, pp. 377–390, Oct./Dec. 2014, doi: https://doi.org/10.1109/TAFFC.2014.2336244.

[13] M. K. Keutmann, S. L. Moore, A. Savitt, and R. C. Gur, "Generating an item pool for translational social cognition research: Methodology and initial validation," Behav. Res. Methods, vol. 47, no. 1, pp. 228–234, Mar. 2015, doi: https://doi.org/10.3758/s13428-014-0464-0.

[14] S. Lee, S. Yildirim, A. Kazemzadeh, and S. S. Narayanan, "An articulatory study of emotional speech production," in Proc. INTERSPEECH 2005, Lisbon, Portugal, Sep. 4–8, 2005, pp. 497–500. Accessed: Feb. 8, 2021. [Online.] Available: https://sail.usc.edu/ema_web/LeeInterSpeech2005.pdf

[15] B. Nojavanasghari, T. Baltrušaitis, C. E. Hughes, and L.-P. Morency, "EmoReact: A multimodal approach and dataset for recognizing emotional responses in children," in Proc. 18th ACM Int. Conf. Multimodal Interaction, Tokyo, Japan, Nov. 12–16 2016, pp. 137–144. doi: https://doi.org/10.1145/2993148.2993168.

[16] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE '05 Audio-Visual Emotion Database," in Proc. 22nd Int. Conf. on Data Eng. Workshops, Atlanta, GA, USA, Apr. 3–7, 2006, p. 8. doi: https://doi.org/10.1109/ICDEW.2006.145.

[17] J. James, L. Tian, and C. Watson, "An open source emotional speech corpus for human robot interaction applications," in Proc. INTERSPEECH 2018, Hyderabad, India, Sep. 2–6, 2018, pp. 2768–2772. doi: https://doi.org/10.21437/Interspeech.2018-1349.

[18] S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, T.-H. Huang, and L.-W. Ku, "EmotionLines: An emotion corpus of multi-party conversations," 2018, arXiv:1802.08379v2. Accessed: Mar. 4, 2021. [Online]. Available: https://arxiv.org/pdf/1802.08379.pdf

[19] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," 2018, arXiv:1810.02508v6. Accessed: Mar. 4, 2021. [Online]. Available: https://arxiv.org/pdf/1810.02508.pdf

[20] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," PLoS ONE, vol. 13, no. 5, p. e0196391, May 16, 2018, doi: https://doi.org/10.1371/journal.pone.0196391.

[21] S. Haq and P. J. B. Jackson, "Multimodal emotion recognition," in Machine Audition: Principles, Algorithms and Systems, W. Wang, Ed., Hershey, PA, USA: IGI Global Press, 2011, pp. 398–423. doi: https://doi.org/10.4018/978-1-61520-919-4.ch017.

[22] M. K. Pichora-Fuller and K. Dupuis, Toronto Emotional Speech Set (TESS). V1. 2020. Distributed by Scholars Portal Dataverse. Accessed: Feb. 8, 2021. doi: https://doi.org/10.5683/SP2/E8H2MF.

[23] N. Holz, P. Larrouy-Maestri, and D. Poeppel, The Variably Intense Vocalizations of Affect and Emotion Corpus (VIVAE). V1. Oct. 5, 2020. Distributed by Zenodo. Accessed: Feb. 8, 2021. [Dataset]. doi: https://doi.org/10.5281/zenodo.4066235.

[24] I. Dzafic, Example emotion videos used in investigation of emotion perception in schizophrenia. 2017. Distributed by the University of Queensland. Accessed: Mar. 3, 2021. [Online]. doi: https://doi.org/10.14264/uql.2017.120.
[25] S. Klaylat, Arabic Natural Audio Dataset Automatic Emotion Recognition. V11. Dec. 1, 2017. Distributed by Kaggle. Accessed: Feb. 8, 2021. [Online]. Available: https://www.kaggle.com/suso172/arabic-natural-audio-dataset/version/11
[26] H. Pajupuu, Eesti Emotsionaalse Kõne Korpus. V5. Jun. 12, 2012. Distributed by Center of Estonian Language Resources. Accessed: Feb. 9, 2021. [Online]. doi: https://doi.org/10.15155/EKI.000A.
[27] L. Kerkeni, C. Cleder, Y. Serrestou, and K. Raoff, French Emotional Speech Database - Oréau. V2. Dec. 31, 2020. Distributed by Zenodo. Accessed: Feb. 9, 2021. [Dataset]. doi: https://doi.org/10.5281/zenodo.4405783.
[28] O. Lahaie and P. Gournay, Canadian French Emotional Speech Database. V1.1. 2017. Distributed by Groupe de Recherche sur la Parole et l'Audio. Accessed: Feb. 8, 2021. [Online]. Available: https://www.gel.usherbrooke.ca/audio/cafe.htm
[29] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in Proc. INTERSPEECH 2005, Lisbon, Portugal, Sep. 4–8, 2005. Accessed: Feb. 9, 2021. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.130.8506&rep=rep1&type=pdf
[30] N. Vryzas, R. Kotsakis, A. Liatsou, C. A. Dimoulas, and G. Kalliris, "Speech emotion recognition for performance interaction," J. Audio Eng. Soc., vol. 66, no. 6, pp. 457–467, Jun. 2018, doi: https://doi.org/10.17743/jaes.2018.0036.
[31] N. Vryzas, M. Matsiola, R. Kotsakis, C. A. Dimoulas, and G. Kalliris, "Subjective evaluation of a speech emotion recognition interaction framework," in Proc. Audio Mostly 2018 Sound Immersion and Emotion, North Wales, United Kingdom, Sep. 12–14, 2018, p. 34. doi: https://doi.org/10.1145/3243274.3243294.
[32] O. M. Nezami, P. J. Lou, and M. Karami, "ShEMO: A large-scale validated database for Persian speech emotion detection," Lang. Resour. and Eval., vol. 53, no. 1, pp. 1–16, Oct. 8, 2018, doi: https://doi.org/10.1007/s10579-018-9427-x.
[33] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "BAUM-1: A spontaneous audio-visual face database of affective and mental states," IEEE Trans. Affect. Comput., vol. 8, no. 3, pp. 300–313, Jul./Sep. 2017, doi: https://doi.org/10.1109/TAFFC.2016.2553038.
[34] C. E. Erdem, C. Turan, and Z. Aydin, "BAUM-2: A multilingual audio-visual affective face database," Multimedia Tools and Applications, vol. 74, no. 18, pp. 7429–7459, May 9, 2015, doi: https://doi.org/10.1007/s11042-014-1986-2.
[35] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," 2020, arXiv:2010.14794v2. Accessed Mar. 3, 2021. [Online]. Available: https://arxiv.org/pdf/2010.14794.pdf
[36] A. Adigwe, N. Tits, K. El Haddad, S. Ostadabbas, and T. Dutoit, "The Emotional Voices Database: Towards controlling the emotion dimension in voice generation systems," 2018, arXiv:1806.09514. Accessed: Feb. 8, 2021. [Online]. Available: https://arxiv.org/pdf/1806.09514.pdf

[37] M. Noordewier and S. Breugelmans, “On the valence of surprise,” Cognition and Emotion, vol. 27, no. 7, pp. 1326–1334, Apr. 2013, doi: https://doi.org/10.1080/02699931.2013.777660.
[38] A. Malek. "SER-datasets." Github. https://github.com/SuperKogito/SER-datasets (accessed Mar. 4, 2021).