Tulane University
SCHOOL OF SCIENCE AND ENGINEERING

Sentiment Analysis of Multimodal Data The Future of Voice Messaging Systems

Andrea Burns, Chloe Chen, Mackenna Barker
Parisa Kordjamshidi
Department of Computer Science

Introduction

Sentiment analysis is an ever growing field of natural language processing (NLP), and it can be expanded with the use of multimodal data in order to classify sentiments in a wider range of emotions (rather than simply positive, negative, or neutral). Sentiment analysis can be applied to a voice messaging system using both the audio (voice recording) and text (transcription) data; our model determines the emotion and urgency of a voice message before a user even hears it.

Problem statement

The overarching goal of our capstone is to create a multimodal machine learning model that is capable of learning and recognizing both the class of emotion and urgency of a voice message. While the primary mode of data will be audio recordings, we also are interested in creating a system that incorporates text data, such as potential transcriptions of the voice message (various developers such as Apple are in process of developing this functionality). Future applications may incorporate text messages.

Given a set of audio and text data, classify the data into one of four emotions – joy, anger, fear, or sadness – and then calculate the intensity of the emotion(s) exhibited.

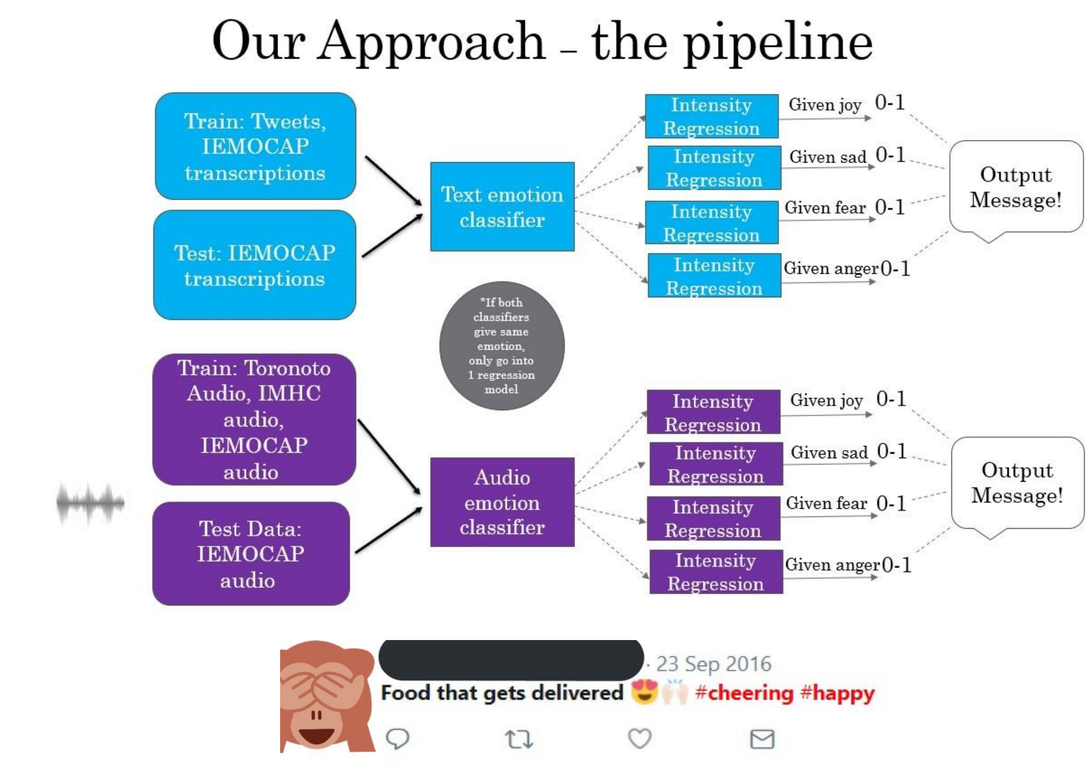


Fig. 1 Pipeline of the project. *Note: if both classifiers return same emotion, only funnel into one regression model

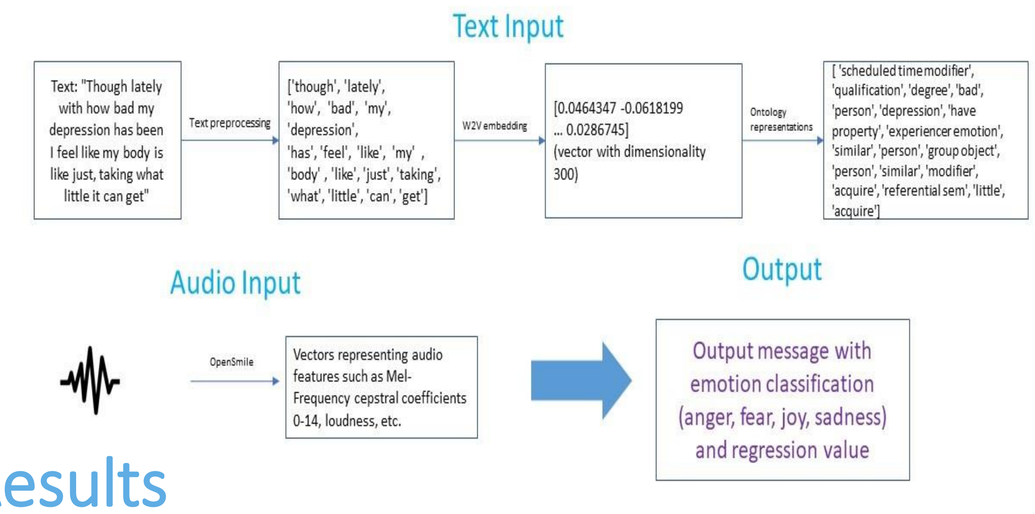
Methods/approach

- Train with text data to build robust system (SemEval-2018 Task 1: Affect in Tweets, 80% of IEMOCAP transcriptions), preprocess using natural language toolkit
- Extract audio features from available datasets (Berlin Database of Emotional Speech, Toronto Emotional Speech Set, 80% of IEMOCAP Dataset) for training classifier
- Classify audio and corresponding transcriptions based on four emotions with multiclass SVM (remaining 20% of IEMOCAP for test)
  - Text Features: Word2Vec word embeddings and ontology representations averaged
  - Audio Features: 1581 features from The INTERSPEECH 2010 Paralinguistic Challenge feature set such as Mel-Frequency cepstral coefficients 0-14, loudness, and more
- Determine intensity using SVM regressor:
  - If both text and audio classifiers return same emotion, funnel into one regression model; otherwise filter data into each respective regression model

References

Poria, Soujanya, et al. “Fusing Audio, Visual and Textual Clues for Sentiment Analysis from Multimodal Content.” *Neurocomputing*, vol. 174, 2016, pp. 50–59.
International Workshop on Semantic Evaluation 2018, “Task 1: Affect in Tweets” http://alt.qcri.org/semeval2018/
C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, December 2008.
Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter Sendlmeier und Benjamin Weiss
A Database of German Emotional Speech
Proceedings Interspeech 2005, Lissabon, Portugal

Input/Output



Results

Performance metrics before combined model:

Emotion classification

	F-1 score*
Extra Trees w/W2V embeddings	0.21
Naïve Bayes w/W2V embeddings	0.26
SVM w/W2V embeddings	0.29
SVM w/W2V embeddings + ontology representations	0.42 (0.97 precision score for 'Fear')

Key Observations: The SVM classifier outperformed the other multiclass classification algorithms consistently, so we chose this model for the final classification algorithm.

Regression values for intensity

	Mean squared error*	Pearson Coefficient	P-value*
Anger Multilayer Neural Net	.032	.077	.33
Anger SVM Regressor	.032	.128	.103
Fear Multilayer Neural Net	.037	.005	.942
Fear SVM Regressor	.037	.162	.02
Joy Multilayer Neural Net	.045	.044	.610
Joy SVM Regressor	.045	.391	.0000025
Sadness Multilayer NN	.035	-.111	.185
Sadness SVM Regressor	.035	.147	.08

Key Observations: SVM consistently performs better in terms of the **Pearson Coefficien**t, which shows the strength of the linear relationship between the predicted intensity scores and the true intensity scores. We strictly use SVMs in the later model due to this advantage. **Mean Squared Error** is as the most popular evaluation metric used in regression problems, which gives a representation of the plausible magnitude of error term. **P-values** indicates how strongly the data contradicts the hypothesis in question.

Final Results for Complete Model

Text Classifier Accuracy: 38.8%

Audio Classifier Accuracy: 40.6%

Percent the classifiers agree: 3.4%

Example output:
This sample was found have anger or sadness with intensity [0.49622346], [0.5036959] respectively.
This sample was found to be anger with intensity [0.50443911]

Observations: All regression values tend to range from .48 to .52, which means they are not capturing emotion well, because these values are neither extreme. The regression model had the least training due to limitations on labeled data, which is most likely the cause of this. Accuracy is as expected, but it appears the classifiers identify emotion differently the majority of the time, showing the complexity of identifying emotion using different modalities.

The text classifier misclassified fear, joy, and sadness as anger and the audio classifier misclassified anger, fear and joy as sadness, joy being of the greatest concern. This is most likely due to unbalanced class samples.

No metrics can be computed for the regression values because we did not have intensity scores for the audio and transcriptions made available by the IEMOCAP dataset.

Conclusions/Discussion/Future work

Multimodal data is intrinsic to human learning, and has recently been acknowledged as a unique tool in creating more robust machine learning classifiers. Previous research set benchmarks for textual-audio-visual multimodality, but our project demonstrated that reasonable accuracy compared to published results can be attained with two-fold data, which will be novel in everyday applications such as voice message systems.