Multimodal Classification of Emotions in Latin Music

Presented by: Damian Domela s1853767

Universiteit Leiden
The Netherlands

2020 IEEE International Symposium on Multimedia

Authors:
Leonardo G. Catharin
Rafael P. Ribeiro
Carlos N. Silla Jr.

Problem Statement

- This paper refers to/builds onto the paper which I used as inspiration for my API project [1]

- Streaming services still struggle with automatically classifying their large music repositories.[1]

- Classification of songs by emotion
    - ‘What emotion is felt by the listener’

The Latin Music Mood Database

- LMMD [2] contains 3,139 audio clips
    - 2,282 unique songs after pre-processing

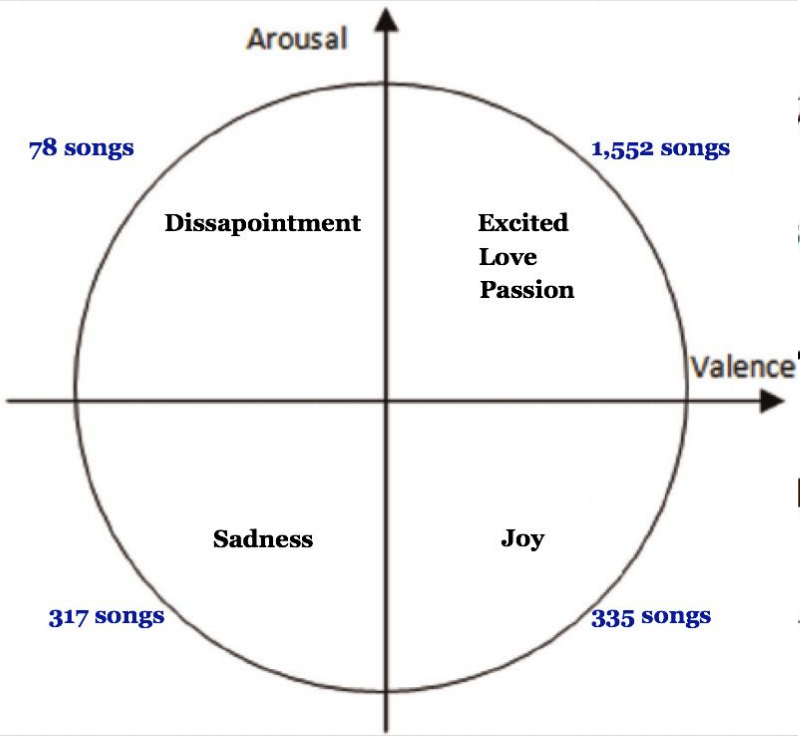
- ‘Ethnic Lyrics Fetcher tool’ is used to retrieve the lyrics for LMMD

- Emotion labels based on Watson’s model [3]
    - Mapping to Russell’s model [4]

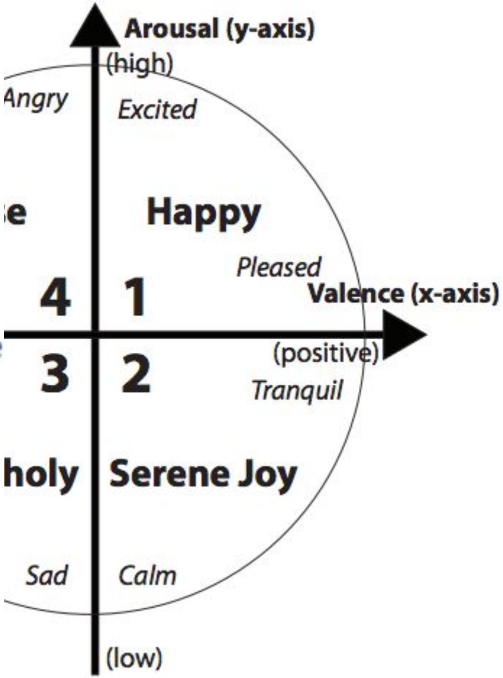
The Latin Music Mood Database

Table 1 Number of songs of each e

	Joy	Sadness	Love	Disap
Axé	53	8	41	8
Bachata	4	59	123	2
Bolero	15	57	114	1
Forró	48	19	88	11
Gaúcha	105	34	30	3
Merengue	27	49	64	6
Pagode	56	30	98	5
Salsa	25	60	90	3
Sertanejo	26	66	67	3
Tango	78	88	60	55



Watson's



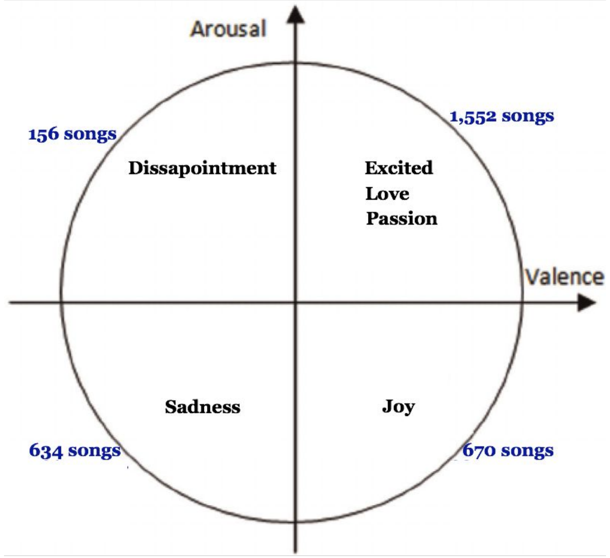
Russell's

Data Imbalance

- May lead to loss of accuracy on minority classes

- SMOTE was used to oversample the minority classes
  - Creating synthetic data points based on existing ones
  - Synthetic data points based on features

- Oversampled Result:



Feature Extraction

Features extracted from lyrics	
Feat. (Dim.)	**Description**
EST (16)	Stylistic Features
2G (614) 3G (4,590) 4G (17,171) ST (13,798)	N-grams normalized with TF-IDF.
2G (100) 3G (100) 4G (100) ST (100)	N-grams normalized with TF-IDF and reduced with LSA.
D2V (250)	Paragraph Vector Embeddings.
Features extracted from audio	
Feat. (Dim.)	**Description**
RLBP (59)	Extracted textural features with RLBP.
BA (48)	MFCC, rolloff, spectral centroid, flux and zerocrossings.
SSD (1,668)	Features SSD and complementary RP e RH.

- Stylistic Features: word counts, lines and special characters
- TF-IDF with Latent Semantic Analysis reduction
- D2V: Word embeddings as features

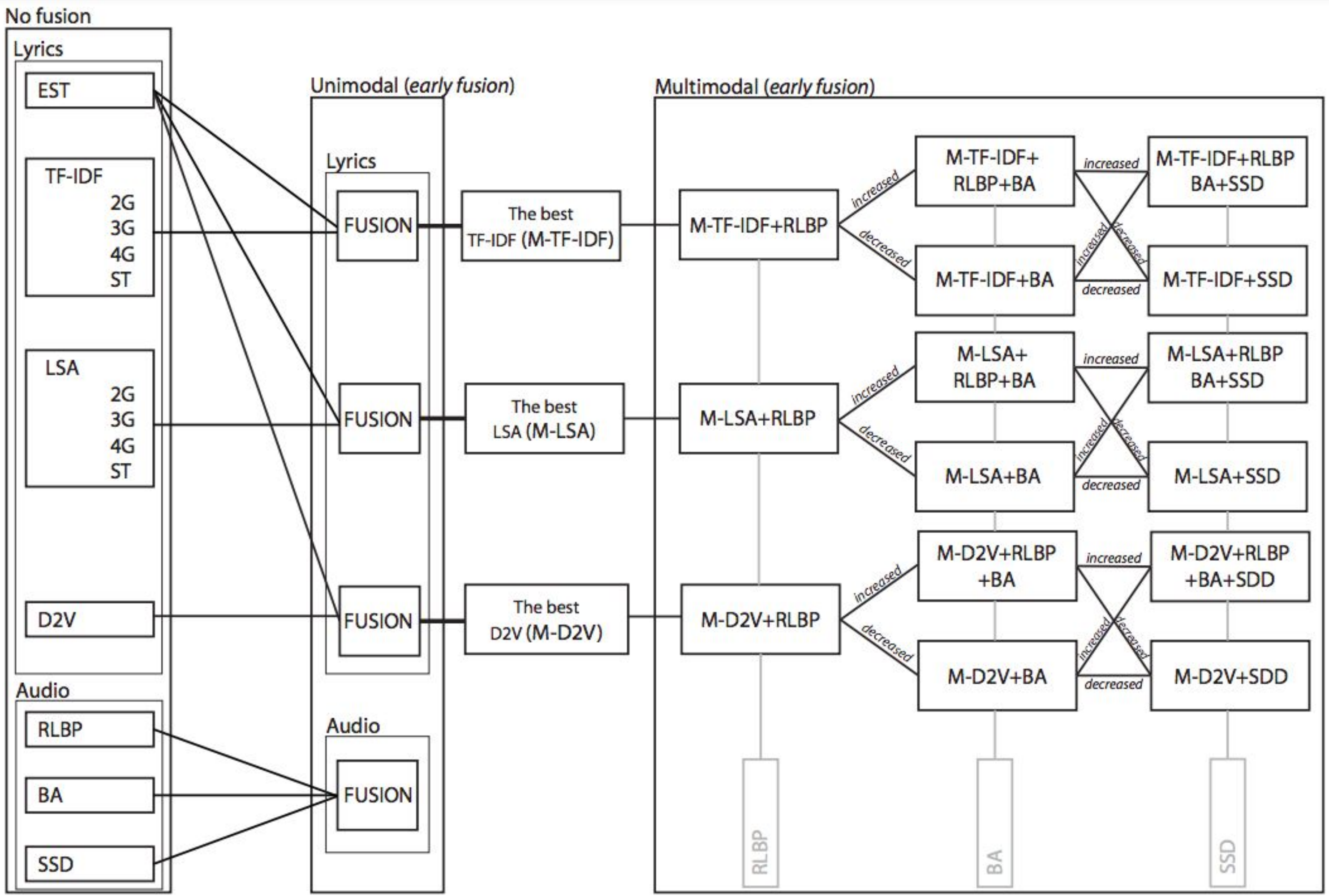
- Robust Local Binary Pattern: Time-frequency spectogram image as input, textural embedded features as output
- Basic Acoustic: Mel-Frequency Cepstral Coefficient, Rolloff, spectral centroid, flux and zerocrossings.
- Statistical Spectrum Descriptor, RP and RH: Directly taken from audio, rhythm fluctuations features

Approach 1: Single-step Classification

- Classify each song with SVM into one of six (Watson's) emotions + map it onto Russell's model.

- Unimodal feature groups: Combined features from the same information source (lyrics/audio)
  - e.g.: Stylistic features + TF-IDF

- Multimodal feature groups: Combines best lyrical unimodal group with each audio feature set
  - e.g.: RLBP + (Stylistic features + TF-IDF)
  - Introduce another audio feature set and assess performance

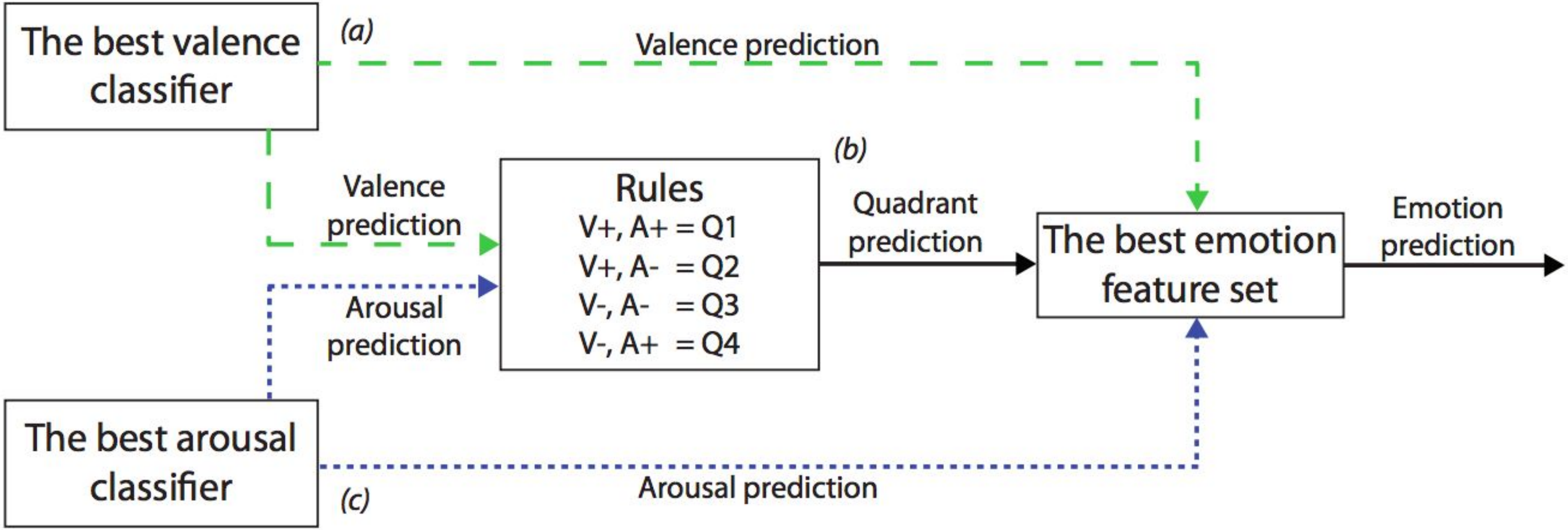

No fusion
Lyrics
EST
TF-IDF
2G
3G
4G
ST
LSA
2G
3G
4G
ST
D2V
Audio
RLBP
BA
SSD
Unimodal (early fusion)
Lyrics
FUSION
FUSION
FUSION
Audio
FUSION
The best TF-IDF (M-TF-IDF)
The best LSA (M-LSA)
The best D2V (M-D2V)
Multimodal (early fusion)
M-TF-IDF+RLBP
M-TF-IDF+RLBP+BA
M-TF-IDF+BA
M-TF-IDF+RLBP BA+SSD
M-TF-IDF+SSD
M-LSA+RLBP
M-LSA+RLBP+BA
M-LSA+BA
M-LSA+RLBP BA+SSD
M-LSA+SSD
M-D2V+RLBP
M-D2V+RLBP +BA
M-D2V+BA
M-D2V+RLBP +BA+SDD
M-D2V+SDD
increased
decreased
RLBP
BA
SSD

Approach 2: Multi-step Classification

- Use prediction of the best valence/arousal Single-step Q classifiers to:
    - Classify corresponding quadrant

- Use valence/arousal/quadrant predictions as features to classify emotions

- The best emotion feature set: feature set from Single-step Q with the best performance.


The best valence classifier
(a)
Valence prediction
Valence prediction
(b)
Rules
V+, A+ = Q1
V+, A- = Q2
V-, A- = Q3
V-, A+ = Q4
Quadrant prediction
The best emotion feature set
Emotion prediction
Arousal prediction
The best arousal classifier
(c)
Arousal prediction

Results Single-step Q

RESULTS FOR SINGLE-STEP VALENCE, AROUSAL, QUADRANT, AND EMOTION CLASSIFICATION WITH AND WITHOUT SMOTE

Feat.		Mean
Val	D2V+EST+RLBP+SSD	0.796
	D2V+EST+RLBP+SSD - SMOTE	0.716
Aro	TF-IDF(4G)	0.700
	TF-IDF(4G) - SMOTE	0.914
Qua	TF-IDF(3G)+SSD	$\boxed{0.644}$
	TF-IDF(3G)+SSD - SMOTE	$\boxed{0.656}$
Emo	TF-IDF(3G)+EST+RLBP+BA	0.470
	TF-IDF(3G)+EST+RLBP+BA - SMOTE	0.481

Results Multi-step Q

MULTISTEP QUADRANT CLASSIFICATION RESULTS

Feat.	**Q1**	**Q2**	**Q3**	**Q4**	**Mean**
Multi-Q	0.890	0.553	0.288	0.215	**0.734**
BestSingle-Q	0.771	0.440	0.343	0.614	**0.656**

Conclusion

- Novel contribution of Emotion Recognition in Spanish/Portuguese context
- State of the art lyrical/audio feature set combinations
- Extremely detailed experiment/results section
- Relatively superior performance to related work

References

[1] Tan, K and Villarino, M and Maderazo, Christian, “*Automatic music mood recognition using Russell’s twodimensional valence-arousal space from audio and lyrical data as classified using SVM and Naïve Bayes*”, IOP Conference Series: Materials Science and Engineering 2019

[2] Carolina L. dos Santos and Carlos N. Silla Jr, “*The Latin Music Mood Database*”, EURASIP Journal on Audio, Speech, and Music Processing (2015)

[3]] D. Watson and A. Tellegen, “Toward a consensual structure of mood.” Psychological bulletin, vol. 98, no. 2, p. 219, 1985.

[4] R. Malheiro, R. Panda, P. Gomes, and R. P. Paiva, “Emotionally-relevant features for classification and regression of music lyrics,” IEEE Transactions on Affective Computing, vol. 9, no. 2, pp. 240–254, 2016.