

Oregon State University

Evaluating Large-Language Models

for Dimensional Music Emotion Prediction from Social Media Discourse

Patrick Donnelly and Aidan Beery

December 17, 2022

soundbendor lab

Introduction

The **task** of **music emotion prediction** attempts to automatically estimate average human affective responses to music.

Introduction

The **task** of **music emotion prediction** attempts to automatically estimate average human affective responses to music.

This **problem** is difficult because of the **paucity of datasets** annotated with subjective human responses.

Introduction

The **task** of **music emotion prediction** attempts to automatically estimate average human affective responses to music.

This **problem** is difficult because of the **paucity of datasets** annotated with subjective human responses.

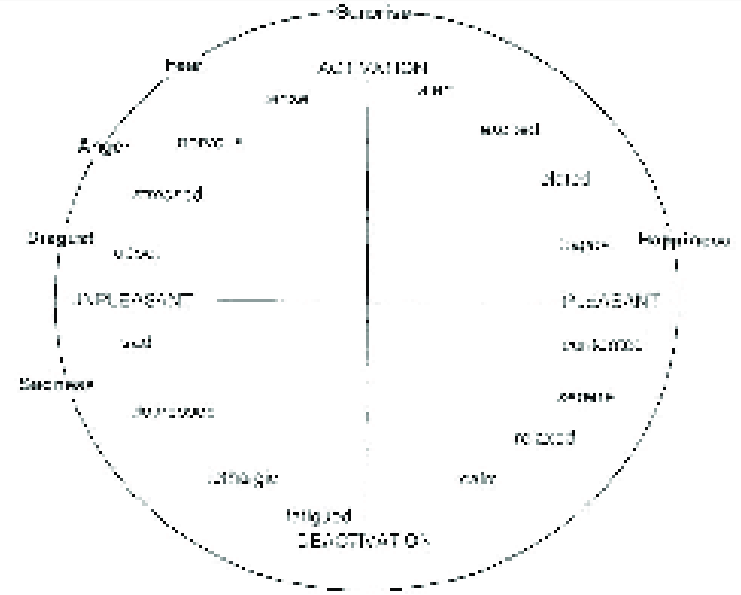
Our **goal** is to **estimate emotion directly from social media** without use of acoustic analysis of the song itself.

Emotion Models

- Plutchik (1980) [1]
  - discrete
  - *joy*, *sadness*, *anger*, *fear*, *trust*, *disgust*, *surprise*, *anticipation*
- Ekman (1992) [2]
  - discrete
  - *anger*, *disgust*, *happiness*, *sadness*, *fear*, *surprise*
- Russell (1980) [3]
  - continuous
  - *valence*, *arousal*

Emotion Models

- Plutchik (1980) [1]
  - discrete
  - *joy, sadness, anger, fear, trust, disgust, surprise, anticipation*
- Ekman (1992) [2]
  - discrete
  - *anger, disgust, happiness, sadness, fear, surprise*
- ***Russell** (1980) [3]
  - continuous
  - *valence, arousal*



Circumplex model (from [3])

Music Emotion Prediction

- Acoustic features [4]
- Song lyrics
  - adding lyrics [5]
    - 89.8% → 92.4%
  - using only lyrics [6]
    - 64.7% for 18 moods
  - using Transformers [7]
    - 94.8% for 4 moods

Related Work

Music Emotion Prediction

- Acoustic features [4]
- Song lyrics
  - adding lyrics [5]
    - 89.8% → 92.4%
  - using only lyrics [6]
    - 64.7% for 18 moods
  - using Transformers [7]
    - 94.8% for 4 moods

Transformers (2017) [8]

- natural language processing
- self-attention for context
- **B**i-directional **E**ncoder **R**epresentations from **T**ransformers (**BERT**) [9]
  - RoBERTa (2019) [10]
  - DistilBERT (2019) [11]

Datasets

Paucity of Data

- no single problem definition
- subjectivity of responses
- expensive lab studies
- copyright concerns

Datasets

Paucity of Data

- no single problem definition
- subjectivity of responses
- expensive lab studies
- copyright concerns

Valence & Arousal

- DEAM [12], n=1803, obscure
- Deezer [13], n=18k, synthetic
- AMG1608 [14], n=1608
- PmEmo [15], n=794

Datasets

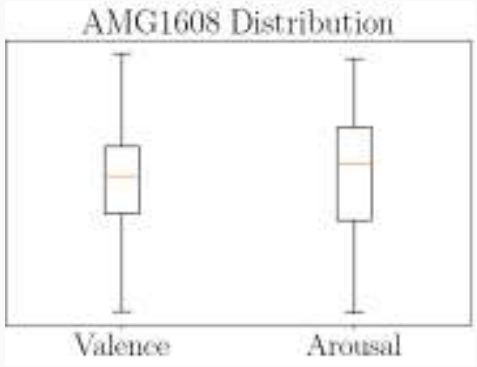
Paucity of Data

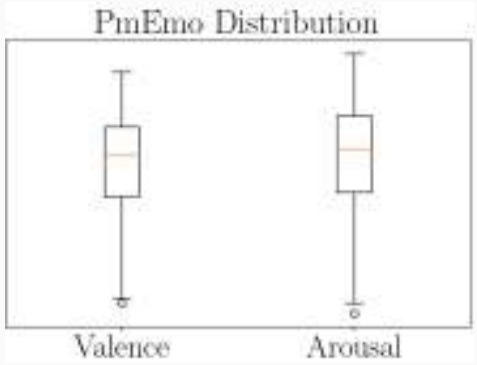
- no single problem definition
- subjectivity of responses
- expensive lab studies
- copyright concerns

Valence & Arousal

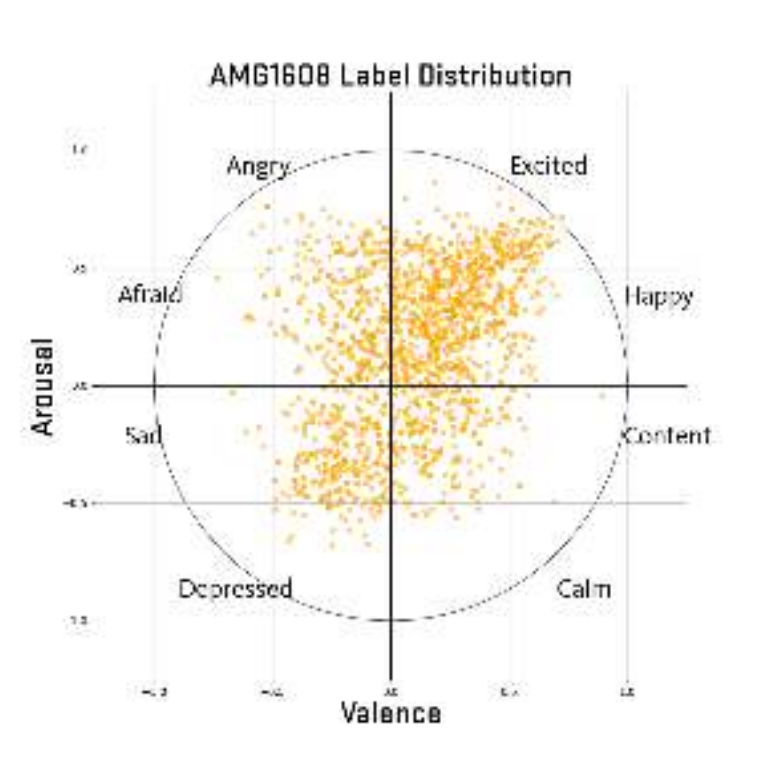
- ~~DEAM [12], n=1803, obscure~~
- ~~Deezer [13], n=18k, synthetic~~
- **AMG1608** [14], n=1608
- **PmEmo** [15], n=794

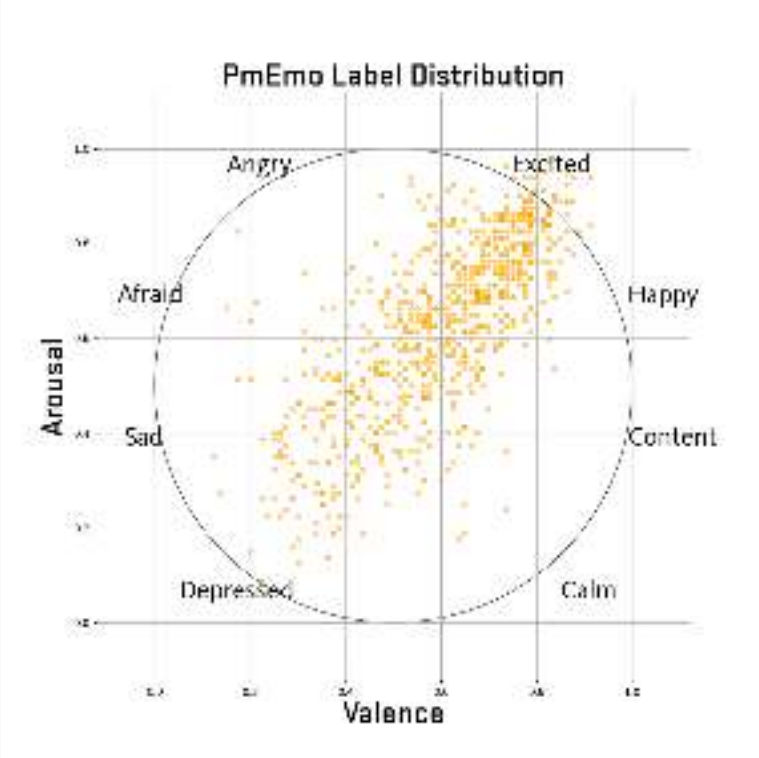
Dataset	Songs	Type	Scaling
AMG1608	1608	Crowd	$[-1, 1]$
PmEmo	794	Lab	$[\ 0, 1\ ]$





Dataset Distribution





Distribution of annotations for AMG1608 and PmEmo.

Social Media Data Collection

- AMG1608 and PmEmo
- two months (Fall 2021)
- query artist and song title
  - excluded if no comments

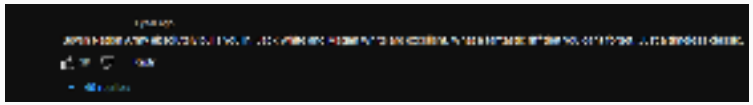
- Youtube and Reddit
  - 10 best comments
  - include nested comments
- Twitter
  - 100 top tweets
  - exclude retweets

		Songs		Comments			Words	
		n	yield	n	μ	σ	μ	σ
AMG1608	Reddit	1431	89%	129,722	80.7	154.3	2400.8	69.1
	YouTube	1592	99%	217,093	135.0	57.7	2128.7	33.66
	Twitter	822	51%	5726	3.6	7.9	51.1	14.5
PmEmo	Reddit	627	86%	103,398	136.6	218.7	3810.5	56.8
	YouTube	730	95%	121,546	160.6	63.9	2172.1	44.1
	Twitter	331	43%	2699	3.6	7.3	46.0	15.2

Social Media Example



Youtube



Reddit

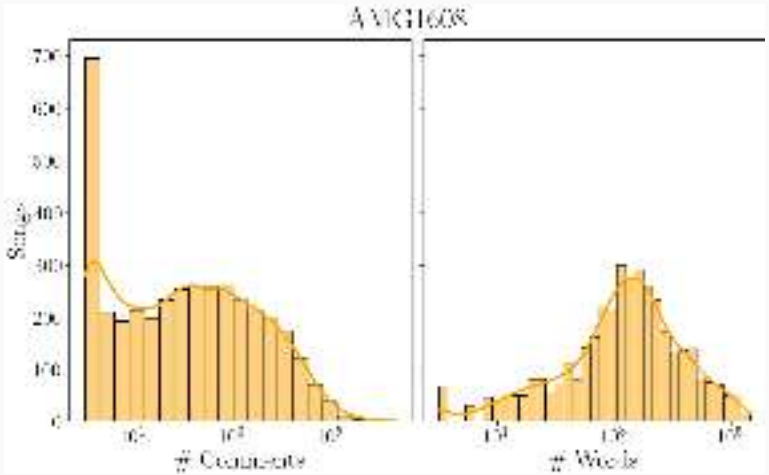


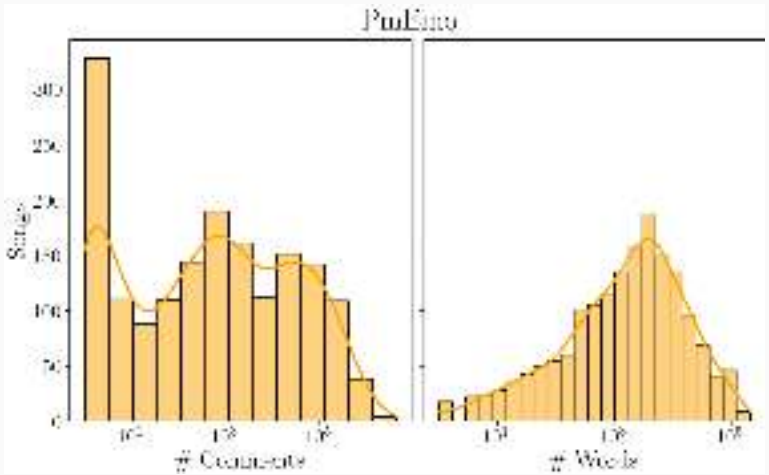
Dataset: **PmEmo**
Artist: **The White Stripes**
Song: **Seven Nation Army**
Valence: **0.5750**
Arousal: **0.6875**

Twitter



Distribution of Comments





Comment and word distributions of our discourse datasets.

Model Comparison

- Model comparison
  - `TFDistilBertBase`
  - `TFRobertaBase`
- default architecture
- 128 word input
- append two dense layers
- MSE loss function
- 70/15/15 dataset split
- predict valence and arousal
- aggregate at song level

		DistilBERT	RoBERTa
AMG1608	Valence	0.49	0.51
	Arousal	0.64	0.63
PmEmo	Valence	0.72	0.71
	Arousal	0.64	0.64

Comparison of DistilBERT and RoBERTa after two epochs of training.

Results by Social Media Source

AMG1608

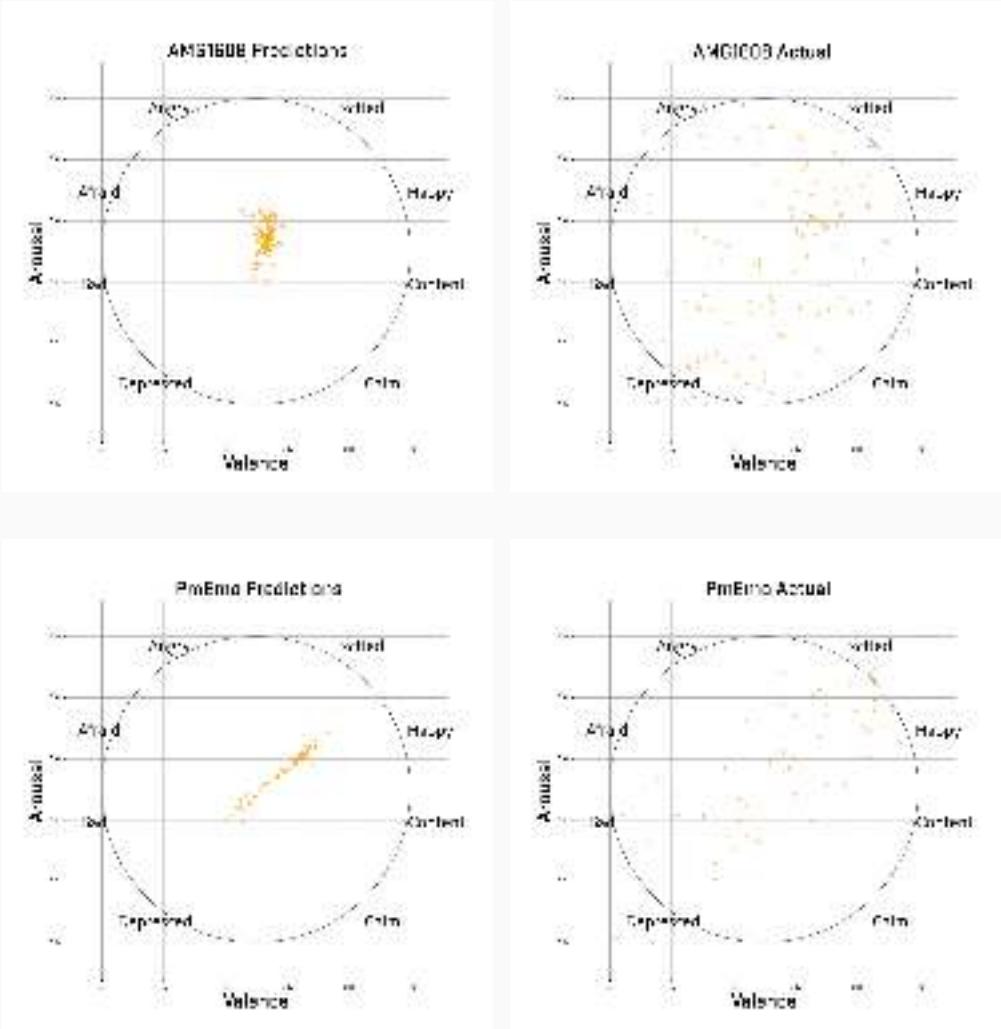
	Reddit	Twitter	YouTube	**All**
Valence	0.32	0.23	**0.62**	0.49
Arousal	0.56	0.34	**0.72**	0.64

PMEmo

	Reddit	Twitter	YouTube	**All**
Valence	0.56	0.26	0.68	**0.72**
Arousal	0.60	0.16	0.52	**0.66**

Results of DistilBERT trained for two epochs for each source and song list.

Limitation: Clustered Predictions



Distribution of DistilBERT predictions for AMG1608 and PmEmo.

Limitations and Future Work

Limitations

Predictions are clustered

1. noisy data;
   lack of emotion terms
2. aggregation at song level

Future Work

1. prune comments
2. other architectures
   - Relation Aware Transformers [16]
   - `xl-net` [17]

Limitations and Future Work

Limitations

Predictions are clustered

1. noisy data;
   lack of emotion terms
2. aggregation at song level

Limited data

- fewer Twitter comments
- Twitter comments are short
- search "all-time" comments

Future Work

1. prune comments
2. other architectures
   - Relation Aware Transformers [16]
   - `xl-net` [17]

Pull additional data

- revise Twitter approach
- LastFM "shouts"
- SoundCloud, time-stamped

Conclusion

Summary

We trained natural language transformer models using *only discourse* from three social media sites to predict the valence and arousal of songs.

- Youtube had highest retrieval rate
- no single social media source dominated
- DistilBERT faster than RoBERTa
- moderate correlation but low accuracy

Conclusion

Summary

We trained natural language transformer models using *only discourse* from three social media sites to predict the valence and arousal of songs.

- Youtube had highest retrieval rate
- no single social media source dominated
- DistilBERT faster than RoBERTa
- moderate correlation but low accuracy

Contribution

These results imply that social media discourse may carry semantic information relevant to predicting human affective responses to music.

References i

[1] R. Plutchik, "A general psychoevolutionary theory of emotion," en, in *Theories of Emotion*, R. Plutchik and H. Kellerman, Eds. Academic Press, Jan. 1980, pp. 3–33, ISBN: 978-0-12-558701-3. DOI: 10.1016/B978-0-12-558701-3.50007-7.

[2] P. Ekman, "An argument for basic emotions," en, *Cognition and Emotion*, vol. 6, no. 3–4, pp. 169–200, May 1992, ISSN: 0269-9931, 1464-0600. DOI: 10.1080/02699939208411068.

[3] J. A. Russell, "A circumplex model of affect.," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980, ISSN: 0022-3514. DOI: 10.1037/h0077714.

[4] R. Panda, R. M. Malheiro, and R. P. Paiva, "Audio features for music emotion recognition: A survey," *IEEE Transactions on Affective Computing*, pp. 1–20, 2020. DOI: 10.1109/TAFFC.2020.3032373.

[5] K. Bischoff, C. S. Firan, R. Paiu, W. Nejdl, C. Laurier, and M. Sordo, "Music mood and theme classification - a hybrid approach," in *Proceedings of the 10th International Society for Music Information Retrieval Conference*, ser. ISMIR, 2009, pp. 657–662. DOI: 10.5281/zenodo.1417317.

[6] E. Cano and M. Morisio, "Moodylyrics: A sentiment annotated lyrics dataset," in *Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, ACM, Mar. 2017, pp. 118–124, ISBN: 978-1-4503-4798-3. DOI: 10.1145/3059336.3059340.

[7] Y. Agrawal, R. G. R. Shanker, and V. Alluri, "Transformer-based approach towards music emotion recognition from lyrics.", vol. 12657, pp. 167–175, 2021. DOI: 10.1007/978-3-030-72240-1_12.

[8] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, 2017, pp. 6000–6010, ISBN: 9781510860964. DOI: 10.48550/arXiv.1706.03762.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *"Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies"*, vol. 1, May 2019, pp. 4171–4186. DOI: 10.48550/arXiv.1810.04805.

[10] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta: A robustly optimized BERT pretraining approach," *arXiv:1907.11692*, vol. 1907, no. 1907.11692, Jul. 2019. DOI: 10.48550/arXiv.1907.11692.

[11] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter," *arXiv:1910.01108*, Oct. 2019. [Online]. Available: https://arxiv.org/abs/1910.01108v4.

References ii

[12] A. Aljanaki and M. Soleymani, "A data-driven approach to mid-level perceptual musical feature modeling," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*, ser. ISMIR, Jun. 2018, pp. 615–621. DOI: 10.48550/arXiv.1806.04903.

[13] R. Delbouys, R. Hennequin, F. Piccoli, J. Royo-Letelier, and M. Moussallam, "Music mood detection based on audio and lyrics with deep neural net," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*, ser. ISMIR, 2018, pp. 370–375. DOI: 10.48550/arXiv.1809.07276.

[14] Y.-A. Chen, Y.-H. Yang, J.-C. Wang, and H. Chen, "The AMG1608 dataset for music emotion recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Apr. 2015, pp. 693–697, ISBN: 978-1-4673-6997-8. DOI: 10.1109/ICASSP.2015.7178058.

[15] K. Zhang, H. Zhang, S. Li, C. Yang, and L. Sun, "The PMEmo dataset for music emotion recognition," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, ACM, Jun. 2018, pp. 135–142, ISBN: 978-1-4503-5046-4. DOI: 10.1145/3206025.3206037.

[16] B. Wang, R. Shin, X. Liu, O. Polozov, and M. Richardson, "Rat-sql: Relation-aware schema encoding and linking for text-to-sql parsers," *arXiv:1911.04942*, Aug. 2021. [Online]. Available: http://arxiv.org/abs/1911.04942.

[17] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, Jan. 2020. DOI: 10.48550/arXiv.1906.08237.