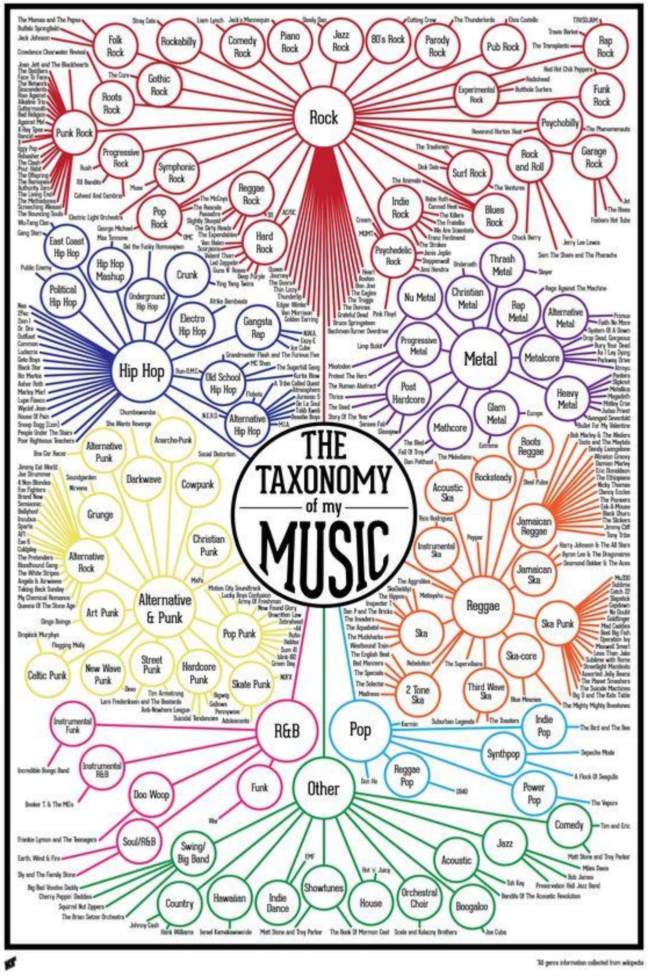
On Combining Diverse Models for Lyrics-Based Music Genre Classification

Caio L. R. S. Ueno
Diego Furtado Silva
Laboratório de Mineração de Dados e Aplicações (MIDAS)
Departamento de Computação
Universidade Federal de São Carlos

Introduction

Automatic music organization and retrieval is a highly required task nowadays, especially for music and video streaming platforms, such as Spotify and YouTube.

However, classifying music by genre is a hard task even for human beings.



Experimental Setup

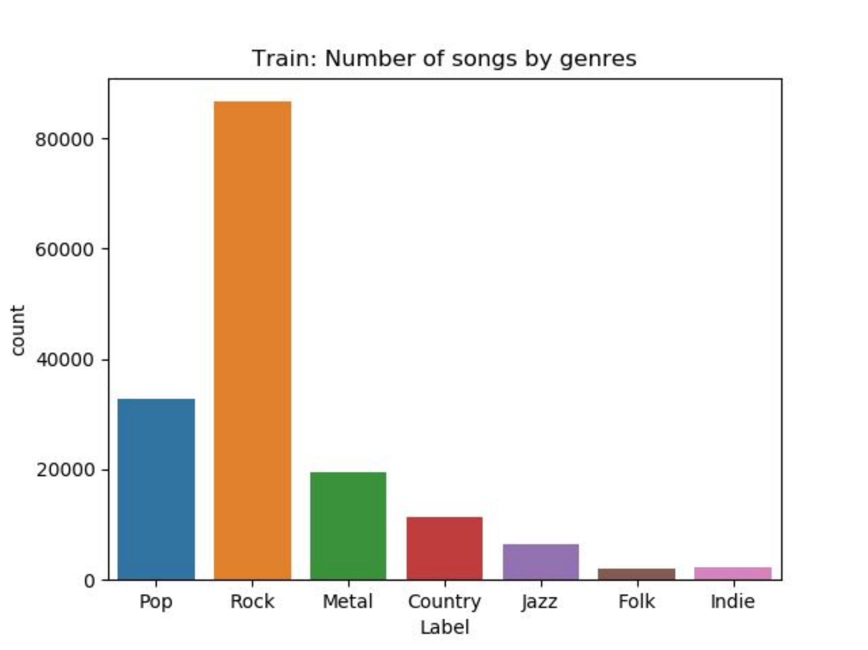
Only the lyrics were used as input for the genre classification models implemented. Some models are based on traditional NLP techniques, essentially bag-of-words, but also on neural networks and word-embedding.

Related work

Tsaptinos published a paper in which he uses a deep neural network lyric-based. Specifically, the author uses a hierarchical attention network (HAN), which considers the text structure, term/sentence/document.

Other papers present different goals, commonly associated with emotion/sentiment analysis. Besides, they also use the respective audio as one of the inputs for their models.

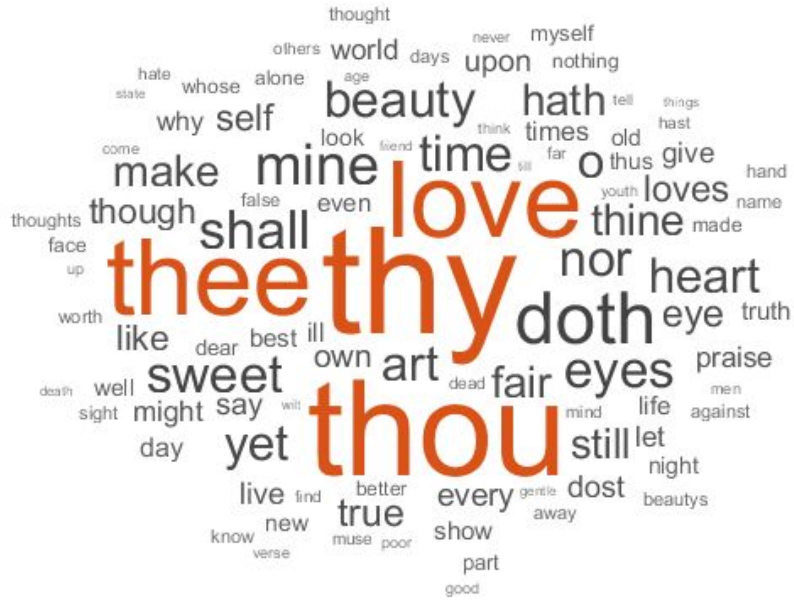
Datasets



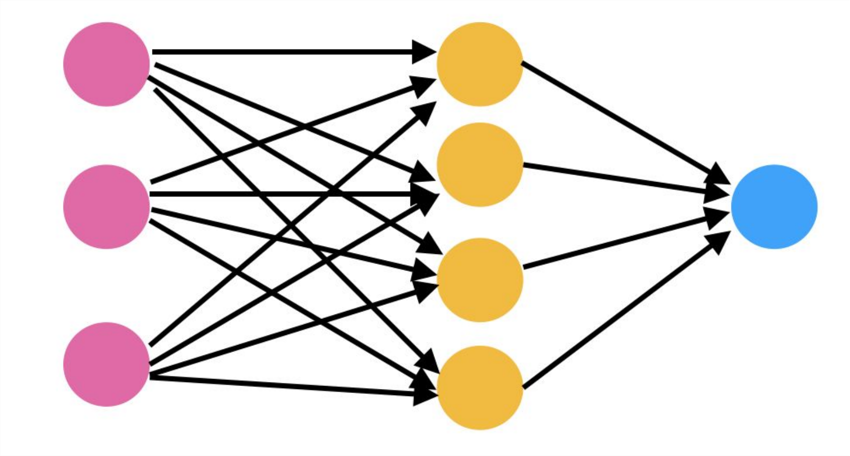
All models were trained and tested with 5 datasets. Only one of them, the smallest, is genre balanced, and it is also a subset of the biggest dataset.

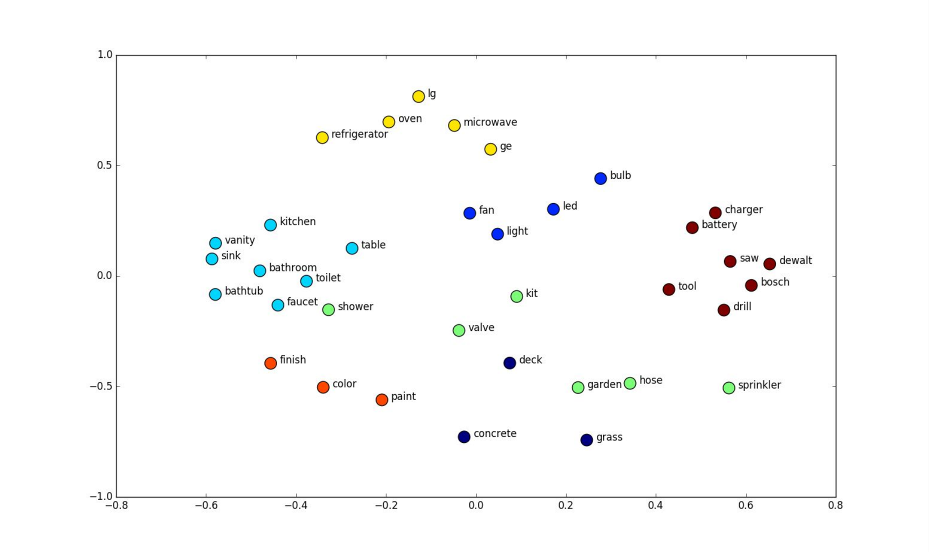
Traditional approach (bag-of-words)

- Naive Bayes
- Support Vector Machines
- Linear Regression
- XGBoosting
- **Random Forest**

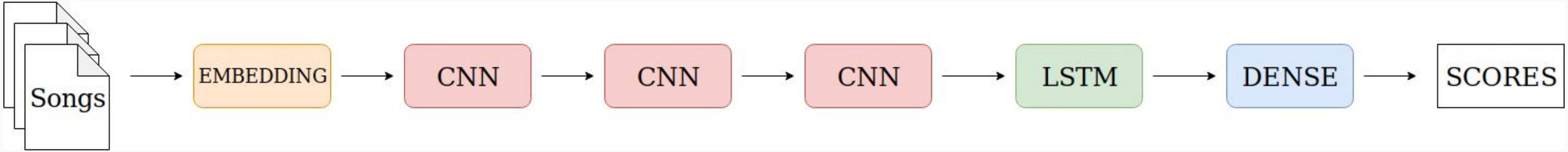


Neural networks and word-embedding

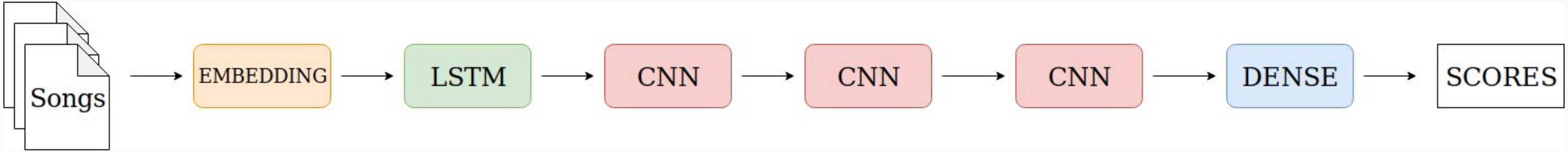




Architectures

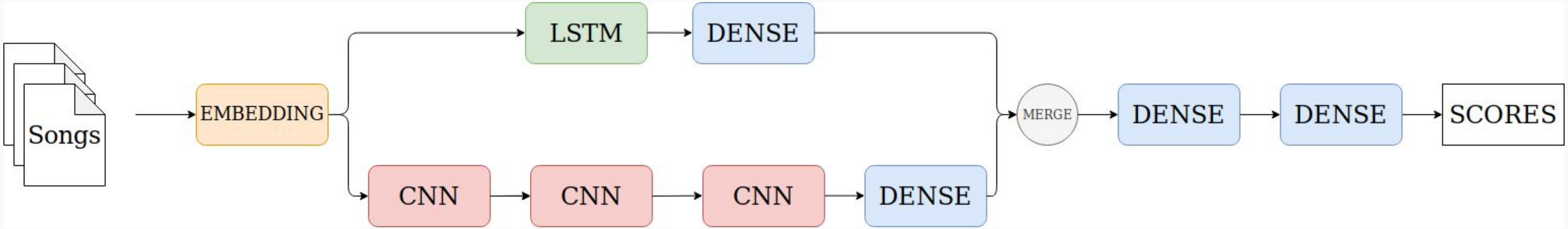


Neural network CNN + LSTM.

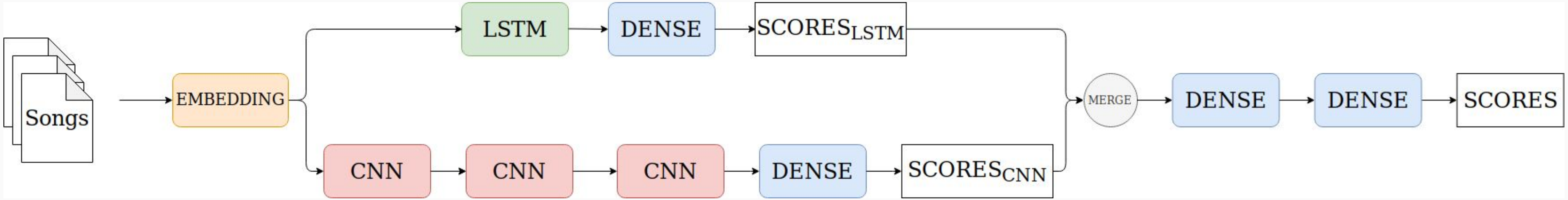


Neural network LSTM + CNN.

Merge

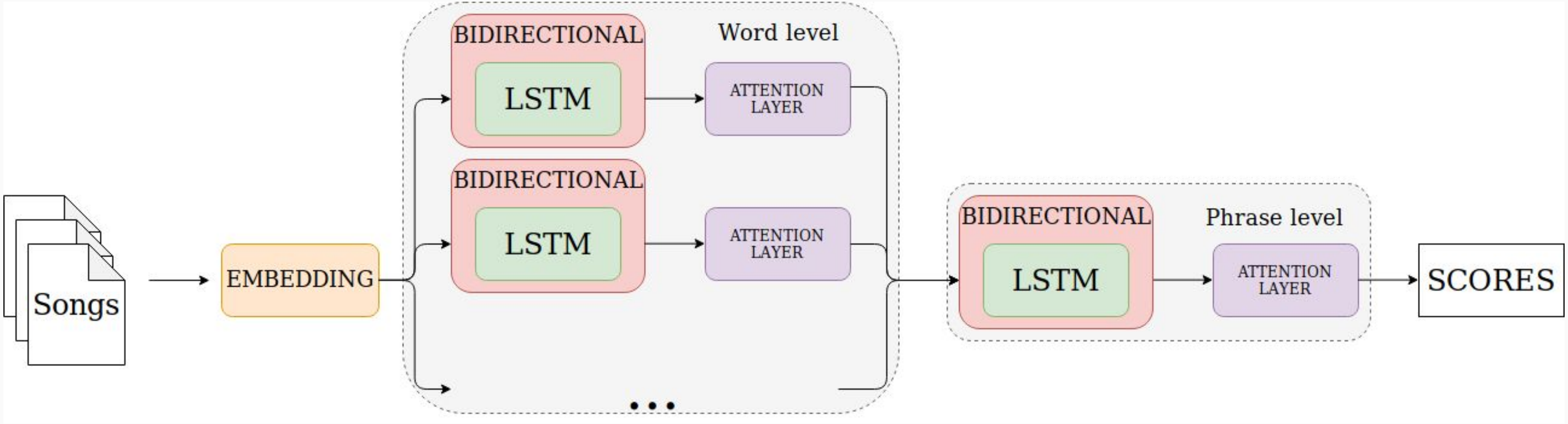


Merge architecture, first variation.



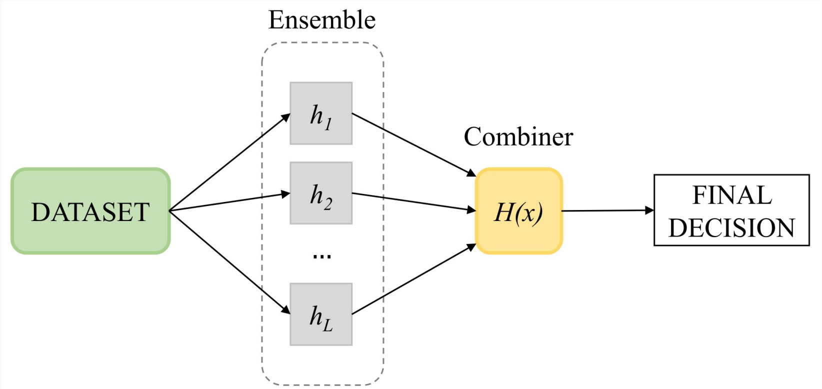
Merge architecture, second variation.

Hierarchical Attention Network



HAN architecture.

Ensembles



Combiner options:

- Major voting
- Weighted voting per classifier
- Score averaging per instance
- Meta-learning

Individual Results

Algorithm	Dataset				
	D1	D2	D3	D4	D5
BoW	**0.646**	**0.608**	**0.604**	**0.523**	**0.434**
CNN	0.582	0.526	0.529	0.422	0.416
LSTM	0.609	0.563	0.586	0.238	0.350
LSTM & CNN	0.593	0.534	0.537	0.512	0.348
CNN & LSTM	0.540	0.479	0.458	0.484	0.136
Merge LSTM&CNN	0.610	0.554	0.534	0.513	0.364
Merge LSTM&CNN 2	0.623	0.582	0.569	0.502	0.330
HAN	0.589	0.568	0.554	0.465	0.413
Best individual accuracy	**0.646**	**0.608**	**0.604**	**0.523**	**0.434**

Ensemble results

It shows that combinations which use both bag-of-words and neural networks based on word-embedding approaches improve the accuracy.

Combined Algorithms	Ensemble Function			
	MV	WV	SA	CL
CNN + LSTM	0.583	0.609	0.634	0.605
CNN + LSTM + HAN	0.637	0.642	**0.650**	0.578
Merge LSTM&CNN + HAN	0.559	0.610	0.638	0.553
BoW + HAN	0.601	0.646	0.639	0.619
BoW + CNN + LSTM	0.610	**0.649**	0.624	**0.674**
BoW + Merge LSTM&CNN	0.573	0.646	0.606	0.631
BoW + CNN + LSTM + HAN	0.623	**0.658**	0.634	0.618
BoW + Merge LSTM&CNN + HAN	0.614	**0.667**	0.629	0.614
All classifiers	**0.668**	**0.669**	**0.673**	0.616

D1 has best individual accuracy equals to 0.646.

Ensemble results

Combined Algorithms	Ensemble Function			
	MV	WV	SA	CL
CNN + LSTM	0.504	0.563	0.580	0.551
CNN + LSTM + HAN	0.586	0.597	**0.610**	0.573
Merge LSTM&CNN + HAN	0.533	0.568	0.596	0.571
BoW + HAN	0.588	0.608	**0.616**	0.540
BoW + CNN + LSTM	**0.648**	**0.626**	**0.650**	0.544
BoW + Merge LSTM&CNN	**0.626**	0.608	**0.647**	0.509
BoW + CNN + LSTM + HAN	**0.659**	**0.634**	**0.666**	0.566
BoW + Merge LSTM&CNN + HAN	**0.655**	**0.624**	**0.664**	0.555
All classifiers	**0.625**	**0.636**	**0.635**	0.569

D2 has best individual accuracy equals to 0.608.

Combined Algorithms	Ensemble Function			
	MV	WV	SA	CL
CNN + LSTM	0.395	0.416	0.419	0.386
CNN + LSTM + HAN	0.429	0.429	**0.438**	0.414
Merge LSTM&CNN + HAN	0.394	0.413	**0.438**	0.401
BoW + HAN	**0.441**	0.434	0.432	0.297
BoW + CNN + LSTM	**0.436**	**0.437**	**0.436**	0.329
BoW + Merge LSTM&CNN	0.403	0.434	0.426	0.268
BoW + CNN + LSTM + HAN	**0.451**	**0.460**	**0.452**	0.294
BoW + Merge LSTM&CNN + HAN	**0.445**	**0.447**	**0.454**	0.279
All classifiers	**0.455**	**0.458**	**0.458**	0.306

Balanced dataset has best individual accuracy equals to 0.434.

References

- Y. Murthy and S. G. Koolagudi, "Content-based music information retrieval (cb-mir) and its applications toward the music industry: A review," ACM Computing Surveys (CSUR) , vol. 51, no. 3, p. 45, 2018.
- A. Tsaptsinos, "Lyrics-based music genre classification using a hierarchical attention network," in International Society for Music Information Retrieval Conference , 2017, pp. 694–701.
- H. Xue, L. Xue, and F. Su, "Multimodal music mood classification by fusion of audio and lyrics," in International Conference on Multimedia Modeling . Springer, 2015, pp. 26–37.
- M. Huang, W. Rong, T. Arjannikov, N. Jiang, and Z. Xiong, "Bi- modal deep boltzmann machine based musical emotion classification," in International Conference on Artificial Neural Networks. Springer, 2016, pp. 199–207.

Thank you!

My e-mail: caioluiggy@hotmail.com
Advisor's e-mail: diego.fsilva@gmail.com

Departamento de Computação
Universidade Federal de São Carlos

Conselho Nacional de Desenvolvimento Científico e Tecnológico

FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO