Identifying Narrative Contexts in Brazilian Popular Music Lyrics Using Sparse Topic Models: A Comparison Between Human-Based and Machine-Based Classification.

André Dalmora[1], Tiago Tavares[1 2]

[1]School of Electric and Computer Engineering (FEEC) – University of Campinas (Unicamp)
CEP 13083-852 – Av. Albert Einstein, 400 – Campinas – SP – Brazil

[2]Interdisciplinary Nucleus for Sound Studies (NICS) – University of Campinas (Unicamp)
CEP 13083-872 – Rua da Reitoria, 165 – Campinas – SP – Brazil

andre.dalmora@gmail.com

Abstract. *Music lyrics can convey a great part of the meaning in popular songs. Such meaning is important for humans to understand songs as related to typical narratives, such as romantic interests or life stories. This understanding is part of affective aspects that can be used to choose songs to play in particular situations. This paper analyzes the effectiveness of using text mining tools to classify lyrics according to their narrative contexts. For such, we used a vote-based dataset and several machine learning algorithms. Also, we compared the classification results to that of a typical human. Last, we compare the problems of identifying narrative contexts and of identifying lyric valence. Our results indicate that narrative contexts can be identified more consistently than valence. Also, we show that human-based classification typically do not reach a high accuracy, which suggests an upper bound for automatic classification.*

1 Introduction

Songs and their lyrics are cultural elements that are frequently linked to the perception of subjective experiences [1, 2]. They can be embedded into cultural activities such as playing games, working, dancing, storytelling, and fighting [3]. The emotional perception of songs is related to inherited cultural aspects of the listener [4]. Regardless of the musical context, the perception of emotions is linked to a person's personal history [5] and to their cultural background [6].

In contemporary Western popular songs, lyrics often refer to similar subjects, such as "a lost love" or "reflections about life". These subjects can be loosely related to the concept of *archetype*, [7, 8]. Archetypes are recurring behavior patterns that can be observed in several elements of the same domain, such as "hero", "sage", or "damsel in distress" in the literature.

The recurring themes of popular songs can also be interpreted as affect-related [9] classification tags. Under this interpretations, affects are described according to a situation presented to a protagonist (which is often the musician themselves) and their reactions related to that, similarly to Russell and Barrett's idea of *prototypical emotional episodes* [10].

In this work, we use text mining tools to automatically classify music lyrics into categories related to their narrative contexts. For such, we built a dataset containing Brazilian popular music lyrics which were raters voted online according to its context and valence. We approached the problem using a machine learning pipeline in which lyrics are projected into a vector space and then classified using general-purpose algorithms. We experimented with document representations based on sparse topic models [11, 12, 13, 14], which aims to find groups of words that typically appear together in the dataset. Also, we extracted part-of-speech tags for each lyric and used their histogram as features in the classification process.

Additionally, we evaluated the classification accuracy difference related to using valence-related categories [14] instead of narrative-based categories. Valence-related categories are more common than story prototypes in text mining [14, 15, 16, 17], and are more connected to subjective perceptions of the listeners.

In order to compare machine and humans we quantified how much humans disagreed with each other. This was made considering each human rater as a predictor of the most-voted labels and calculating their accuracy. Based on the controversy of the ratings, we analyzed the performance of the machine learning algorithms.

Our results suggest that the effectiveness of machine-learning classification and human-based classification are comparable, yet machines are outperformed by human-based classification. Also, they indicate that classifying narrative contexts is an easier task than classifying valence. Last, results show that machines yield lower accuracy when classifying narrative contexts that are controversial among the human raters.

In addition, we foster scientific reproducibility and continuity by making our dataset available online at `http://www.github.com/aldalmora/NC_Music_Brazil`.

The remainder of this paper is organized as follows. Section 2 describes the methods used in this work. Section 3 discusses the results and their outcome. Section 4 resumes the results and suggests future research about this subject.

2 Methods

In this work, we analyzed the predictive power of text mining tools to classify music lyrics according to their narrative contexts and valence. First, we built a manually-classified ground-truth dataset, as described in Section 2.1. Then, we conducted prediction experiments as discussed in Section 2.2.

2.1 Dataset

The dataset used in this work contains 380 song lyrics extracted from Brazilian Bossa-Nova songs. The lyrics in the dataset were classified according to their narrative contexts. The classification used anonymous raters and an online voting system. Each lyric was classified at least five times. We used the most voted classification as the ground-truth. The narrative context categories used for this classification were *relationships* (songs about a romantic relationships), *emotions or reflections* (songs about abstract impressions on life), and *others* (other themes). Also, the lyrics were classified according to their valence (positive, neutral, or negative) using the same procedure.

The number of songs in each narrative context and valence categories are shown in Table 1. Lyrics that did not have a most voted option (that is, there was a draw between two or more options) within subject or valence were discarded from the dataset.

Table 1: Description for the dataset classified by anonymous rankers showing the amount of lyrics classified according to their narrative context (Relationships, Emotions/Reflections, or Others) and their valence (Positive, Neutral, or Negative).

Context	Pos.	Neu.	Neg.	**Total**
Rel	35	40	29	**138**
Emo/Ref	41	37	10	**123**
Others	10	18	7	**44**
Total	**86**	**95**	**46**	**227**

The next section presents the predictive analysis using text mining and machine-learning algorithms.

2.2 Predictive analysis

In this section, we describe our results related to predictive analysis. We tested several classifier variations, as depicted in Figure 1. Each of the variations used a different representation for the documents.

The first variation regards removal of stopwords. Stopwords are words that are too common in a language and convey little meaning. For this reason, stopwords can be removed from text documents without changing its meaning. However, they can be important for predictive analysis. Hence, we evaluated classifiers both with and without stopword removal.

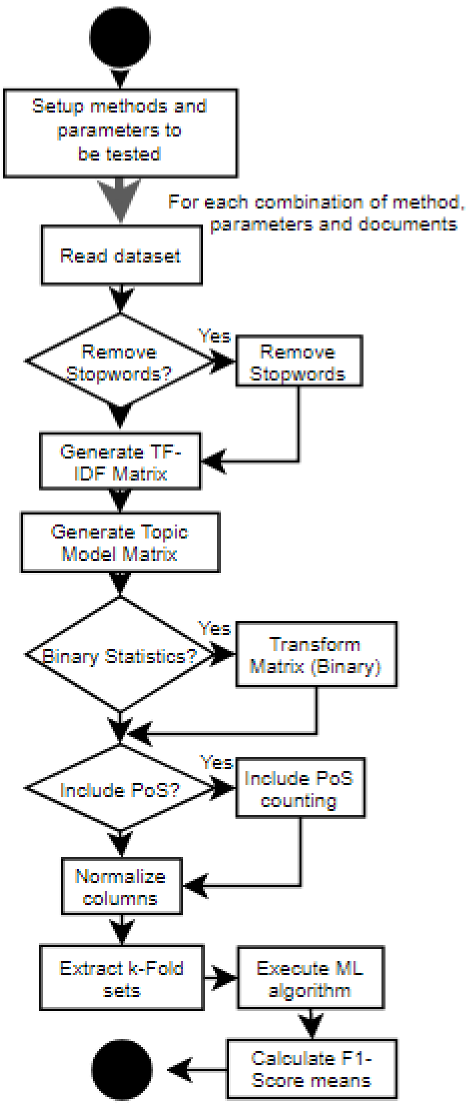


Figure 1: Flowchart depicting the evaluation script for testing diverse configurations and hyper-parameters of the ML algorithms.

After deciding for removing or keeping stopwords, each word in the lyrics is mapped to a vector $l \in \mathbb{R}^P$, where P is the total number of words in the whole dataset and l_p is the TF-IDF [18] rating of word p. The whole set of D lyrics vectors are grouped in a matrix $\boldsymbol{L} \in \mathbb{R}^{D \times P}$. Then, the document matrix $\boldsymbol{L}$ is factorized into two other matrices, as:

$$\boldsymbol{L} \approx \boldsymbol{D}\boldsymbol{W}. \quad (1)$$

In Equation 1, matrix $\boldsymbol{D}$ represents the association of each document to an specific topic, and matrix $\boldsymbol{W}$ represents the association between topics and dataset words. The number of topics n is the inner dimension in the multiplication $\boldsymbol{DW}$, and is a hyperparameter of this process. The values of the elements of $\boldsymbol{D}$ and $\boldsymbol{W}$ are obtained by minimizing:

$$|\boldsymbol{L} - \boldsymbol{DW}|^2 + \lambda|\boldsymbol{D}|, \quad (2)$$

where $\lambda = 0.5$ is a regularization factor used to foster sparsity in D by minimizing the L1-norm $|\boldsymbol{D}|$.

We evaluated two variations of the topic ($\boldsymbol{D}$) representation. The first directly used the results of minimizing Equation 2, meaning the strength of a topic in each lyric. The second one used a binary representation for $\boldsymbol{D}$ in which all non-zero values were mapped to one, meaning the presence or absence of a topic in each lyric.

Another variation regards the use of part-of-speech (PoS) tags. These tags are related to the classification of words as verbs, nouns, and others. They were used by counting the amount of each tag present in each document and including as columns in the representation matrix $\boldsymbol{D}$.

After building the representation matrix of the documents, we used a K-Folds cross-validation schema for testing the algorithm. We used the mean unweighted F1-Score as the evaluation metric. The F1-Score is calculated as:

$$\text{Precision} = \frac{\text{\# true positives}}{\text{\# true positives} + \text{\# false positives}} \tag{3}$$

$$\text{Recall} = \frac{\text{\# true positives}}{\text{\# true positives} + \text{\# false negatives}} \tag{4}$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

We tested Support-Vector Machines (SVM), K Nearest Neighbors (KNN), Gaussian Naive Bayes (GNB), and Random Forest (RF) classifiers. All feature set combinations were tested with each of the classifiers.

Additionally, we compared the machine-based classification accuracy to the average human classification accuracy. For such, we used three different subsets of the dataset. They respectively comprised song lyrics whose most voted label had 3, 4, and 5 votes. This separation aimed at detecting whether more controversial (from the raters' perspectives) lyrics is related to different machine-learning based classification accuracy.

The results of these tests are discussed in the next section.

3 Results and Discussion

The results presented in this section are divide in three parts. First we analysis the outcome from machine learning predictions. Then, in the second part, we analysis the consistency of the ground-truth used in for training ML models, focusing on the online raters classifications and their divergence. Last we present and discuss the behavior of the ML algorithms related to controversial lyric classifications by humans.

3.1 Machine-Learning Classification

After executing the different methods described in the previous section, we compiled the best results as presented in Table 2 and Table 3. Topics(from topic analysis) indicates how the topic representation was quantified. PoS indicates whether Part-Of-Speech was considered or not. SW indicates if the stopwords were removed from the lyrics. Table 2 contains results generated with the objective of predicting the context. Table 3, with the objective of predicting the valence. The 0R classifier works by selecting the most frequent class in the dataset and was used as a baseline.

Table 2: Accuracy and per-class F1-Scores for predicting the Context.

Method	Topic	PoS	SW	F1	Acc.
KNN	Binary	No	No	0.49 ±0.14	0.56
RF	Count	No	No	0.44 ±0.06	0.62
SVC	Binary	No	No	0.51 ±0.11	0.56
GNB	Count	No	No	0.52 ±0.07	0.55
0R	-	-	-	-	0.45

Table 3: Accuracy and per-class F1-Scores for predicting the Valence.

Method	Topic	PoS	SW	F1	Acc.
KNN	Binary	Yes	No	0.50±0.14	0.55
RF	Binary	No	No	0.40±0.11	0.51
SVC	Binary	Yes	No	0.51±0.10	0.56
GNB	Binary	No	No	0.51±0.13	0.52
0R	-	-	-	-	0.42

The results shown in tables 2 and 3 show that removing stop words does not increase classification accuracy. Also, they indicate that Part-of-Speech tags are only relevant for predicting valence. Last, we note that the Gaussian Naive Bayes classifier had the best performance in both cases.

These results were compared to the accuracy of human raters, as shown in the next section.

3.2 Comparison to Human classification

The dataset used in this work was the result of voting using human raters. This means that raters can disagree with each other. The test described in this section evaluates the rating dispersion of each rater.

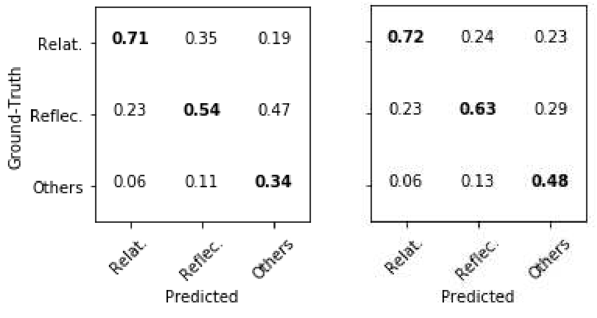
In this test, we consider each rater as a predictor, and calculate their F1-Score. After that, we calculated the average F1-Score for each classification task (predicting context or valence). The results are shown in Table 4.

Table 4: F1-Score for classification performed by humans and machines each dataset rater as a predictor.

Target	0R	Machine	Humans
Context	0.45	0.525±0.07	0.644±0.27
Valence	0.42	0.511±0.13	0.565±0.28

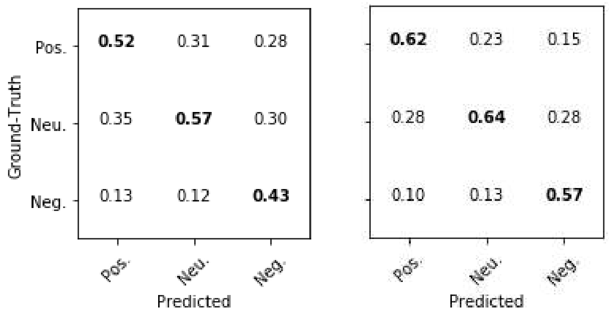
The confusion matrices related to the human and machine classifications are shown in Figures 2 and 3.

We can notice that machine learning was unable to outperform human classification in all objectives and classes. Also, we can notice that the *relationship* class is related to less errors in the narrative context classification.



(a) Machine **(b) Humans**

Figure 2: Confusion matrices for the prediction of narrative contexts.



(a) Machine **(b) Humans**

Figure 3: Confusion matrices for the prediction of valence.

Last, we can see that *neutral* valence was the easiest to classify.

Next, we show how machine learning classifiers behave with more controversial lyrics.

3.3 Behavior with Controversial Lyrics

Some lyrics in our dataset were classified more consistently than others. This consistency can be measured according to the number of human raters that agree with the most frequently voted label. Using this criterion, we can build the following subsets for our dataset:

- $G5$ - Five votes
- $G4$ - Four votes
- $G3$ - Three votes

By grouping the results relate to the best classifier discussed in the previous sections according to these subsets, we obtain the classification accuracy shown in Table 5 for context labelling and Table 6 for valence labeling.

These results show that lyrics that are rated more consistently in narrative contexts can also be classified more accurately by the machine. This suggests that the machine learning algorithms are converging towards learning behaviors that are similar to common sense.

Table 5: Accuracy results for narrative context classification divided by lyric rating divergence.

Method	G5	G4	G3
KNN	0.63	0.56	0.49
RF	0.74	0.63	0.53
SVC	0.58	0.60	0.51
GNB	0.69	0.56	0.51
Human	1.00	0.80	0.60

Table 6: Accuracy results for valence classification divided by lyric rating divergence.

Method	G5	G4	G3
KNN	0.44	0.49	0.48
RF	0.50	0.54	0.47
SVC	0.50	0.54	0.52
GNB	0.49	0.49	0.51
Human	1.00	0.80	0.60

However, this behavior cannot be observed for valence classification. We speculate that valence ratings are more closely related to each rater's personal experiences than to objective elements of text. This reflected on the lower classification ratings for humans as shown in Table 4, and in a lower consistency in the data yielded to machine-learning algorithms.

The next section presents conclusive remarks.

4 Conclusion

This paper described a series of experiments using text mining tools to classify lyrics based on their narrative contexts and their valences. Our results were compared to the ones yielded by human raters using a dataset built specifically for this work.

The results show that classification results are sensibly different when different target labels are used. Also, they show that different document representations can be more effective for the classification of different labels.

Results also show that valence classification is a harder task than narrative context classification. We speculate that this is because valence rating is highly influenced by the rater's personal experience, whereas rating narrative contexts is a more objective task.

Last, results show that narrative contexts that are more controversial for human raters are also harder to classify using machine learning algorithms. This indicates that, in this case, the classification algorithm behaves similarly to the rater's common sense.

This work did not evaluate representations that take phrase structures into account, like attention networks. This poses an interesting avenue for future work.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior – Brasil (CAPES) – Finance Code 001.

References

[1] Thomas Scheff. *What's Love got to do with It? Emotions and Relationships in Popular Songs*. Routledge, 2016.

[2] Yi-Hsuan Yang and Homer H. Chen. Prediction of the distribution of perceived music emotions using discrete samples. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2184 – 2196, 2011.

[3] Andrew H. Gregory. *The Social Psychology of Music*, chapter 7, pages 123–140. Oxford University Press, 1997.

[4] Eduard Hanslick. *On The Musically Beautiful*. Hackett Publishing Company, 1986.

[5] Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1):1–23, 10 2016.

[6] Yukiko Uchida, Sarah S. M. Townsend, Hazel Rose Markus, and Hilary B. Bergsieker. Emotions as within or between people? cultural variation in lay theories of emotion expression and inference. *Personality and Social Psychology Bulletin*, 35(11):1427–1439, 2009. PMID: 19745200.

[7] Joseph Campbell. *The hero with a thousand faces*. New World Library, 2008.

[8] C. G. Jung. *The Archetypes and the Collective Unconscious*. Princeton University Press, 1980.

[9] Paula Dornhofer Paro Costa. *Two-Dimensional Expressive Speech Animation*. PhD dissertation, Universidade Estadual de Campinas - Faculdade de Engenharia Elétrica e de Computação, 2015.

[10] James A Russell and Lisa Feldman Barrett. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, 76(5):805, 1999.

[11] Sanjeev Arora;Rong;Ankur Moitra. Learning topic models - going beyond SVD. *CoRR*, abs/1204.1956, 2012.

[12] C. Jareanpon, W. Kiatjindarat, T. Polhome, and K. Khongkraphan. Automatic lyrics classification system using text mining technique. In *2018 International Workshop on Advanced Image Technology (IWAIT)*, pages 1–4, Jan 2018.

[13] Y. Hu and M. Ogihara. Identifying accuracy of social tags by using clustering representations of song lyrics. In *2012 11th International Conference on Machine Learning and Applications*, volume 1, pages 582–585, Dec 2012.

[14] Swati Chauhan;Prachi Chauhan. Music mood classification based on lyrical analysis of hindi songs using latent dirichlet allocation. *2016 International Conference on Information Technology (InCITe) - The Next Generation IT Summit on the Theme - Internet of Things: Connect your Worlds*, pages 72–76, 2016.

[15] Stuti Shukla;Pooja Khanna;Krishna Kant Agrawal. Review on sentiment analysis on music. *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*, pages 777–780, 2017.

[16] Lili Nurliyana Abdullah Teh Chao Ying, Shyamala Doraisamy. Genre and mood classification using lyric features. *2012 International Conference on Information Retrieval & Knowledge Management*, pages 260–263, 2012.

[17] Felipe S. Tanios;Tiago F. Tavares. Impact of genre in the prediction of perceived emotions in music. *16th Brazilian Symposium on Computer Music*, 2017.

[18] J.D. Rajaraman, A.;Ullman. Data mining. In *Mining of Massive Datasets*, pages 1–17. Cambridge University Press, 2011.