UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

LEONARDO GABIATO CATHARIN

Classificação automática de emoções a partir de letras e áudios de músicas

MARINGÁ

2020

LEONARDO GABIATO CATHARIN

Classificação automática de emoções a partir de letras e áudios de músicas

Dissertação apresentada ao Programa de Pós Graduação em Ciência da Computação do Departamento de Informática, Centro de Tecnologia da Universidade Estadual de Maringá, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Orientadora: Valéria Delisandra Feltrim

Coorientador: Yandre Maldonado e Gomes da Costa

MARINGÁ

2020

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá - PR, Brasil)

C361c

Catharin, Leonardo Gabiato

Classificação automática de emoções a partir de letras e áudios de músicas / Leonardo Gabiato Catharin. -- Maringá, PR, 2020.

117 f.: il. color., figs., tabs.

Orientadora: Profa. Dra. Valéria Delisandra Feltrim.

Coorientador: Prof. Dr. Yandre Maldonado e Gomes da Costa.

Dissertação (Mestrado) - Universidade Estadual de Maringá, Centro de Tecnologia, Departamento de Informática, Programa de Pós-Graduação em Ciência da Computação, 2020.

1. *Latin Music Mood Database* (LMMD). 2. Aprendizagem de máquina. 3. Músicas (Letras e áudios) - Emoções - Classificação automática (Informática). I. Feltrim, Valéria Delisandra, orient. II. Costa, Yandre Maldonado e Gomes da, coorient. III. Universidade Estadual de Maringá. Centro de Tecnologia. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação. IV. Título.

CDD 23.ed. 006.45

Elaine Cristina Soares Lira - CRB-9/1202

FOLHA DE APROVAÇÃO

LEONARDO GABIATO CATHARIN

Classificação automática de emoções a partir de letras e áudios de músicas

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Departamento de Informática, Centro de Tecnologia da Universidade Estadual de Maringá, como requisito parcial para obtenção do título de Mestre em Ciência da Computação pela Banca Examinadora composta pelos membros:

BANCA EXAMINADORA

Profa. Dra. Valéria Delisandra Feltrim
Universidade Estadual de Maringá – DIN/UEM

Prof. Dr. Yandre Maldonado e Gomes da Costa
Universidade Estadual de Maringá – DIN/UEM

Prof. Dr. Marcos Aurélio Domingues
Universidade Estadual de Maringá – DIN/UEM

Prof. Dr. Juliano Henrique Foleis
Universidade Tecnológica Federal do Paraná – DACOM/UTFPR-CM

Aprovada em: 31 de julho de 2020.
Local da defesa: Sala virtual
https://meet.google.com/ege-hyuq-bhf .

AGRADECIMENTOS

A Deus, por me abençoar com saúde e inteligência para atingir meus objetivos.

À minha família, por sempre me apoiar em minhas decisões e mais do que isso, me dar a oportunidade de estudar e chegar até aqui.

À UEM e a todos os meus professores, por me darem a oportunidade de realizar e concluir o mestrado.

À professora Dra. Valéria Delisandra Feltrim e ao Prof. Dr. Yandre Maldonado e Gomes da Costa, por me orientarem e incentivarem na realização deste trabalho.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) pelo apoio financeiro.

A todos os meus amigos que participaram do meu dia a dia e me ajudaram a concluir essa fase da minha vida.

Classificação automática de emoções a partir de letras e áudios de músicas

RESUMO

As músicas fazem parte do cotidiano das pessoas. Além de serem uma forma de comunicação, elas são utilizadas como forma de entretenimento, ou ainda em tratamentos médicos e terapias. Como forma de comunicação, pode-se dizer que as músicas podem expressar emoções. Além disso, acredita-se que a busca de músicas pelas pessoas está relacionada ao que elas estão sentindo naquele momento. Surge, assim, o desafio de oferecer músicas identificadas por emoções a usuários de conteúdo musical. Isso é um problema especialmente para os provedores de música, já que há uma elevada quantidade de músicas nos repositórios e a tarefa de rotular essas músicas por emoções é cara e complexa. Dessa forma, se torna necessário criar mecanismos automáticos para tal tarefa. Existem diversos trabalhos na literatura que abordam esse problema. Na maioria deles, letras ou áudios das músicas são utilizados como fonte de informação para a extração de características, que, por sua vez, alimentam algoritmos de aprendizagem de máquina. Muitos desses trabalhos são voltados para a língua inglesa e poucos tratam músicas latinas. Visando preencher essa lacuna, este trabalho teve como objetivo classificar as músicas da *Latin Music Mood Database* (LMMD) por emoções utilizando características extraídas das letras e áudios. As características foram avaliadas individualmente, por combinação unimodal (características da mesma fonte de informação) e por combinação multimodal (características das duas fontes de informação). A partir das características extraídas, três abordagens de classificação foram propostas, a saber: classificação simples, na qual as músicas foram classificadas diretamente em emoções; classificação em etapas, na qual as músicas foram classificadas primeiramente em valência, excitação e quadrante e, posteriormente, em emoções; e classificação por *ensemble*, na qual as músicas foram classificadas diretamente em emoções utilizando os *ensemble Adaboost* e *Bagging*. Como a LMMD é desbalanceada em relação às emoções e isso afetava o desempenho dos classificadores, a técnica *Synthetic Minority Over-sampling Technique* (SMOTE) foi aplicada. Os experimentos mostraram que a classificação em etapas com a aplicação da técnica SMOTE obteve melhores resultados em relação às outras abordagens.

Palavras-chave: Letras. Áudios. Emoções. Músicas. LMMD. Classificação automática.

Automatic emotion classfication based on lyrics and audio

ABSTRACT

Music is part of people's daily lives. In addition to being a form of communication, they are used as a form of entertainment, or even in medical treatments and therapies. As a form of communication, it can be said that songs express emotions. In addition, it is believed that people's search for music is related to their feelings at that moment. Thus, the challenge arises of offering music identified by emotions to users of musical content. This is a problem especially for music providers, as there are a lot of songs in the repositories and the task of labeling them by emotions is expensive and complex. Thus, it becomes necessary to create automatic mechanisms for this task. There are several works in the literature that address this problem. In most of them, lyrics or audios of the songs are used as a source of information for extracting features that feed machine learning algorithms. Many of these works focus on the English language and only a few deals with Latin music. To fill this gap, we classify the songs in the Latin Music Mood Database (LMMD) by emotions, using features extracted from lyrics and audios. The features were evaluated individually, using a unimodal combination (features of the same information source) and a multimodal combination (features of the two sources of information). From these features, three classification approaches were proposed, namely: simple classification, in which the songs were classified directly into emotions; classification in stages, in which the songs were classified first in valence, excitement and quadrant and, later, into emotions; and classification by ensemble, in which the songs were classified directly into emotions using Adaboost and Bagging. As the LMMD is unbalanced in relation to emotions and this affected the performance of the classifiers, the Synthetic Minority Over-sampling Technique (SMOTE) technique was applied. The experiments showed that the classification in stages with the application of the SMOTE technique obtained better results in relation to the other approaches.

Keywords: Lyrics. Audio. Emotion. Music. LMMD. Automatic classification.

LISTA DE QUADROS

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

SVM	*Support Vector Machine*
MER	Reconhecimento de Músicas por Emoção
AM	Aprendizagem de Máquina
LMMD	*Latin Music Mood Database*
TF-IDF	Frequência dos termos-Inverso da frequência dos documentos
EST	Características estilísticas
LSA	Análise de Semântica Latente
D2V	*paragraph vector*
RLBP	*Robust Local Binary Pattern*
MFCC	*Mel-frequency cepstral coefficients*
SSD	*Statistical Spectrum Descriptor*
RH	*Rhythm Histogram*
RP	*Rhythm Patterns*
SMOTE	*Synthetic Minority Over-sampling Technique*
MIR	Recuperação de Informações de Músicas
BoW	*Bag-of-words*
TF	Frequência dos termos
IDF	Inverso da frequência dos documentos
SVD	Decomposição de Valores Singulares
Word2Vec	*Word vectors*
CBOW	*Continuous bag-of-words*
PV-DM	*Distributed Memory Model of Paragraph Vectors*
LBP	*Local Binary Pattern*
FFT	*Fast Fourier Transform*
ISMIR	*International Society for Music Information Retrieval*
POS	Part of Speech
LIWC	*Linguistic Inquiry and Word Count*
ANEW	*Affective Norms for English Words*
GI	*General Inquirer*
HMM	*Hidden Markov Model*
MLP	*Multilayer perceptron*
CCA	Análise de correlação canônica
SVDD	*Support Vector Domain Description*
FFNN	*Feed-forward Neural Network*
CNN	Rede Neural Convolucional
LPQ	*Local Phase Quantization*
LMD	*Latin Music Database*
LMLMD	*Latin Music Lyrics Mood Database*
SMO	*Sequential minimal optimization*

SUMÁRIO

1 INTRODUÇÃO

Cada vez mais as músicas fazem parte do cotidiano das pessoas. Em muitos casos, as músicas são utilizadas como forma de entretenimento, porém como afirma Tavares et al. (2017), alguns especialistas acreditam que músicas podem ser utilizadas em tratamentos médicos e terapias. A música também é uma forma de comunicação e a maneira como é composta pode expressar diversos aspectos subjetivos, como por exemplo a emoção que o artista pretende transmitir. Assim, pode-se dizer que músicas podem expressar emoções. Por outro lado, acredita-se que a busca de músicas pelas pessoas tem uma forte relação com os sentimentos que elas estão sentindo naquele momento. Nesse contexto, torna-se um desafio para fornecedores de conteúdo musical oferecer músicas que possam ser identificadas por suas emoções.

Como afirma Tan et al. (2019), a indústria da música ainda encara a classificação de músicas por emoções como um problema devido à quantidade de músicas disponíveis e a dificuldade dessa tarefa ser realizada manualmente. É possível encontrar diversas bases musicais que são disponibilizadas por diversos fornecedores. Muitas dessas bases não possuem um sistema com uma filtragem de músicas por emoções e as que possuem, como a allmusic.com[1] e o Spotify[2], recorrem a uma classificação manual ou parcialmente manual. Assim, se torna necessário criar mecanismos para automatizar a tarefa de classificação de músicas por emoção.

Muitos trabalhos podem ser encontrados na literatura visando a solução desse problema. Trabalhos como o de Przybysz (2016) apresentam resultados promissores usando diversas fontes de informações para extração de características que são utilizadas por classificadores baseados em aprendizagem de máquina. Também são comuns na literatura trabalhos que utilizam duas fontes específicas de informação: áudios e letras. Por meio dessas fontes de informação é possível extrair características e usá-las em classificadores supervisionados, como o *Support Vector Machine* (SVM). Por meio do áudio é possível observar aspectos como ritmo, timbre e intensidade da música. Esses aspectos são fontes de características que auxiliam na tarefa de Reconhecimento de Músicas por Emoção (MER). Já as letras ajudam a concentrar a atenção do ouvinte e podem despertar emoções específicas.

[1] https://www.allmusic.com/
[2] https://www.spotify.com/br/

A psicologia, desde 1930, estuda e interpreta o valor afetivo das palavras, com base em levantamentos empíricos (YANG; LEE, 2009). Nesse contexto, o estudo das letras de música para a classificação por emoções das mesmas tem crescido.

Em geral, trabalhos que abordam a tarefa de MER, utilizam Aprendizagem de Máquina (AM) supervisionada e, com isso, se faz necessário o uso de bases de dados rotuladas para os treinamentos e testes dos algoritmos. Para este trabalho, as músicas que foram utilizadas são provenientes da base *Latin Music Mood Database* (LMMD) (SANTOS; SILLA, 2015). Essa base foi escolhida por ser talvez a única base de músicas latinas disponível com rótulo de emoção, letra e áudio disponíveis. A base contém 3.139 áudios de músicas divididas em seis emoções, sendo elas: alegria, amor, decepção, excitado/entusiasmado, paixão e tristeza. Para criar uma versão multimodal dessa base, as letras das músicas foram recuperadas por Ribeiro (2015) que conseguiu 2.603 letras. O *matching* desses áudios e letras foi realizado como parte deste trabalho.

Neste contexto, o foco deste trabalho foi classificar músicas por emoções usando características extraídas das letras e áudio das músicas. Para as letras foram utilizadas características superficiais, como Frequência dos termos-Inverso da frequência dos documentos (TF-IDF) e características estilísticas (EST), e também foram exploradas características que modelassem aspectos semânticos, como Análise de Semântica Latente (LSA) e *paragraph vector* (D2V). Para os áudios foram exploradas características visuais e acústicas. As características visuais foram extraídas de espectrogramas, que são representações particulares de tempo-frequência gerados a partir dos sinais de áudio (FULOP, 2011). Já as características acústicas foram extraídas diretamente do sinal do áudio. Foi utilizado o *Robust Local Binary Pattern* (RLBP) para extrair características visuais, enquanto que foram utilizadas *Mel-frequency cepstral coefficients* (MFCC) combinadas com *Rolloff*, *Spectral centroid*, *Flux* e *Zerocrossings* e os descritores *Statistical Spectrum Descriptor* (SSD), *Rhythm Histogram* (RH) e *Rhythm Patterns* (RP) para extrair as características acústicas. Essas características foram avaliadas em três fases, a saber: fase sem combinação, fase unimodal e fase multimodal. Na fase sem combinação as características foram avaliadas individualmente; na fase unimodal as características de uma mesma fonte de informação (áudio ou letras) foram combinadas; e, por fim, na fase multimodal as melhores características encontradas nas letras foram combinadas com as características extraídas dos áudios.

A partir das características extraídas, três abordagens de classificação foram propostas, a saber: classificação simples, classificação em etapas e classificação por *ensemble*. Na classificação simples as músicas foram classificadas diretamente nas seis emoções da base

LMMD. Nessa fase, o processo de avaliação de características foi aplicado para identificar o conjunto de características de melhor desempenho. Na classificação em etapas, as músicas foram classificadas em valência, excitação e quadrante e, a partir dessas predições, classificadas em emoções. Para encontrar o melhor classificador de valência, excitação e quadrante, o processo de avaliação de características também foi aplicado. Além disso, a combinação das predições dos melhores classificadores de valência e excitação também foi avaliada como classificador de quadrante. Nessas duas primeiras abordagens, os experimentos foram realizados com o classificador *Support Vector Machine* (SVM), já que implementações de SVM se destacaram na revisão da literatura realizada. Foi observado que a base LMMD é desbalanceada e isso afetou negativamente os desempenhos dos classificadores. Para atenuar isso, a técnica *Synthetic Minority Over-sampling Technique* (SMOTE) foi aplicada nos melhores classificadores de emoção, valência, excitação e quadrante. Por fim, na classificação por *ensemble*, as músicas foram classificadas diretamente nas seis emoções da base LMMD utilizando os *ensemble Adaboost* e *Bagging*. Nessa abordagem, as melhores características encontradas na classificação simples foram avaliadas.

Na classificação simples, o melhor resultado obtido foi atingido com uma combinação multimodal. Esse resultado foi superado na classificação em etapas, utilizando a combinação das predições dos melhores classificadores de valência e excitação junto com o melhor conjunto de características encontrado na classificação simples. Dessa forma, a classificação em etapas superou a classificação simples mostrando que, subdividir o problema de classificação pode ser uma estratégia melhor do que classificar diretamente as músicas por emoções. Os resultados da classificação por *ensemble* foram inferiores aos das classificações simples e em etapas. O melhor resultado foi atingido com *Adaboost* e *Random Forest*.

O restante deste trabalho está organizado como segue. Na Seção 2 são apresentados conceitos que fundamentam o tema proposto neste trabalho. Na Seção 3 são apresentados os resultados da revisão sistemática realizada bem como os demais trabalhos relacionados que foram recuperados. Já na Seção 4 é apresentada a metodologia utilizada para o desenvolvimento deste trabalho. Na Seção 5 são apresentados os resultados e discussões dos experimentos realizados. Por fim, na Seção 6 são apresentadas as conclusões e possíveis trabalhos futuros. Após a conclusão são listadas as referências bibliográficas utilizadas e os apêndices do trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

A Recuperação de Informações de Músicas (MIR) é uma área de pesquisa que concentra estudos relacionados aos aspectos de organização e recuperação de coleções musicais (PRZYBYSZ, 2016). Tarefas como a recuperação de músicas por gêneros musicais ou por emoções têm se destacado nessa área. Em especial, o Reconhecimento de Emoções em Músicas (MER) tem recebido uma atenção crescente da comunidade cientifica, já que a emoção é uma característica crucial da música (AN et al., 2017).

Nesse contexto, como será mostrado nas Subseções 3.1 e 3.2, é possível encontrar diversos trabalhos na literatura que propõem métodos automáticos para a tarefa de reconhecimento de emoções em músicas. As letras e os áudios das músicas aparecem como as fontes de informações com melhor desempenho. As características extraídas dessas duas fontes de informação podem tanto ser utilizadas individualmente, quanto de forma combinada.

Na Subseção 2.1 são apresentados conceitos relacionados às emoções e seus aspectos. Na Subseção 2.2 são apresentadas características que podem ser extraídas de letras e áudios das músicas. Na Subseção 2.3 são apresentados o problema de desbalanceamento e fundamentos do algoritmo SMOTE que atenua esse problema. Por fim, na Subseção 2.4 são apresentados conceitos sobre os algoritmos de classificação utilizados neste trabalho.

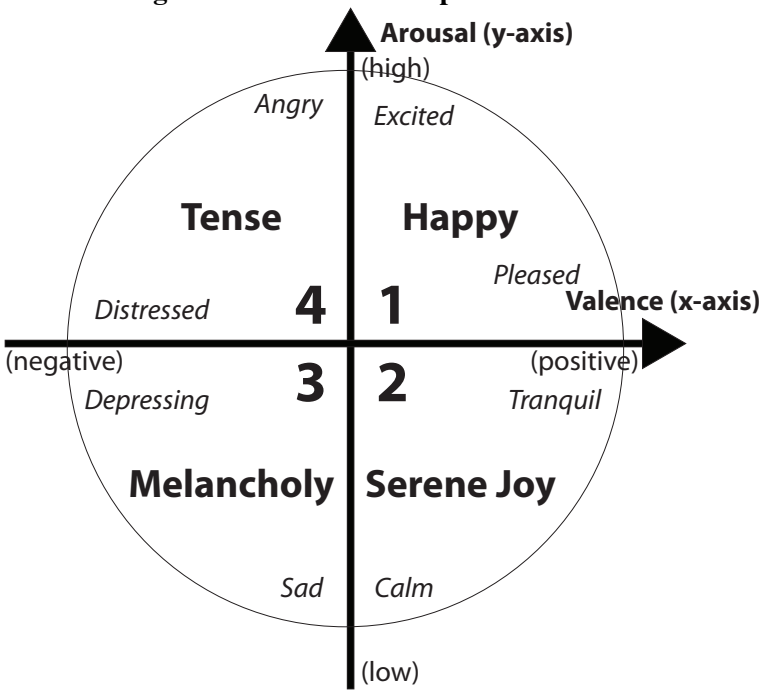
2.1 EMOÇÕES

As emoções são vivenciadas diariamente pelas pessoas. Mesmo assim, é muito difícil expressar como o ser humano reconhece essas emoções por meio de uma teoria científica (RUSSELL, 1980). Apesar disso, alguns pesquisadores, em especial da área da psicologia, propuseram modelos de organização de emoções a fim de relacioná-las e compreendê-las melhor. Nos modelos apresentados neste trabalho, optou-se por não realizar traduções quanto as emoções por causa do risco em não se ter traduções precisas dessas emoções.

Um modelo clássico encontrado na literatura é o modelo circumplexo de Russell (1980). Nesse modelo, as emoções são dispostas em um plano bidimensional em que o eixo horizontal, denominado valência, determina o nível de polaridade da emoção e o eixo vertical, denominado excitação (*arousal*), determina o nível de intensidade da emoção. Na Figura 1 o

modelo de Russell é apresentado.

Figura 1: Modelo circumplexo de Russell



Fonte: Adaptado de Russell (1980)

Por meio da Figura 1 é possível visualizar a disposição das emoções no plano bidimensional e assim ter uma visão organizada dessas emoções. O modelo ainda pode ser interpretado em quatro quadrantes. No quadrante 1 situam-se as emoções com excitação (*Arousal*) e valência (*Valence*) positivas. A emoção *Excited* (animado) é um exemplo desse quadrante. No quadrante 2 encontram-se as emoções com valência positiva, porém com excitação negativa. A emoção *Calm* (calmo) é um exemplo desse quadrante. Já no quadrante 3 encontram-se as emoções com excitação e valência negativas. Um exemplo desse quadrante é a emoção *Sad* (Triste). Por fim, no quadrante 4 encontram-se as emoções com excitação positiva e valência negativa. Tais emoções são negativas, porém geram um alto nível de excitação. A emoção *Angry* (Bravo) é um exemplo desse quadrante.

O modelo de Russell, apesar de intuitivo, tem sido adaptado de diferentes formas na literatura de MER. Patra et al. (2018) basearam-se nesse modelo e criaram uma taxonomia que agrupa as emoções em cinco classes, como mostra o Quadro 1. Cada coluna do quadro representa uma classe de emoções e elas foram agrupadas de acordo com suas semelhanças. Por exemplo, as emoções *Happy*, *Delighted* e *Pleased* foram agrupadas por possuírem alta valência, enquanto que as emoções *Excited*, *Astonished* e *Aroused* foram agrupadas por possuírem alta excitação.

Quadro 1: Taxonomia de emoções proposta por Patra et al. (2018)

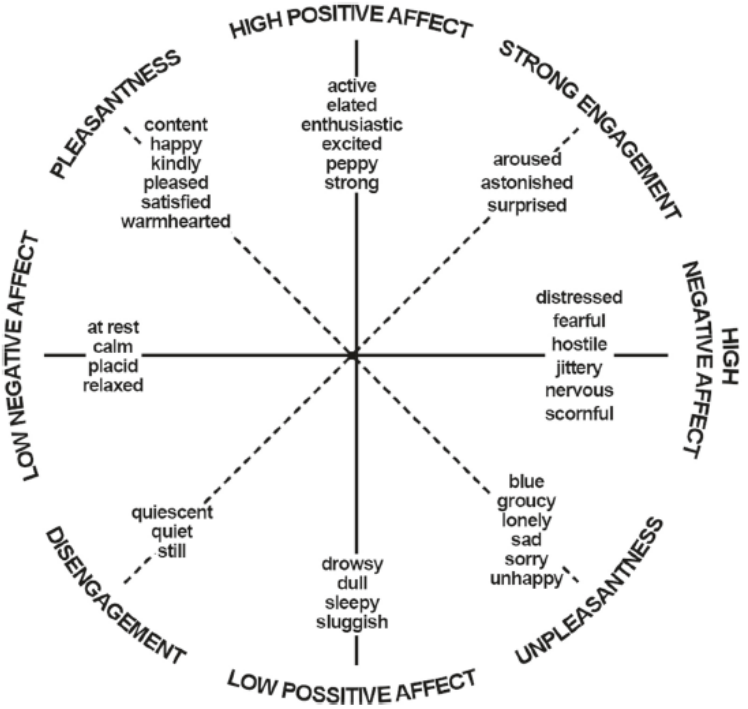
Class_Excited	Class_Happy	Class_Calm	Class_Sad	Class_Angry
Excited	Happy	Calm	Sad	Angry
Astonished	Delighted	Relaxed	Gloomy	Alarm
Aroused	Pleased	Satisfied	Depressed	Tensed

Fonte: Adaptado de Patra et al. (2018)

Malheiro et al. (2018) também utilizaram o modelo de Russell, mas abstraíram as emoções como classes e utilizaram apenas as informações dos quadrantes. Dessa forma, os autores utilizaram os valores de excitação e valência para classificar as músicas nos quatro quadrantes do modelo.

Outro modelo de organização das emoções, também baseado em um plano bidimensional, foi proposto por Watson e Tellegen (1985). Nesse modelo, os eixos do plano representam o afeto positivo (eixo vertical) e o afeto negativo (eixo horizontal). Para construir esse modelo os autores realizaram uma ampla análise de vários modelos de emoções, tal como o de Russell (1980). A Figura 2 apresenta a estrutura consensual de emoções proposta por Watson e Tellegen (1985).

Figura 2: Estrutura consensual de Watson e Tellegen (1985)



Fonte: Watson e Tellegen (1985)

Diferentemente do modelo de Russell (1980), o modelo de Watson e Tellegen (1985)

não estabelece a intensidade de cada emoção. Dessa forma, não é possível afirmar que a emoção "*Happy*" seja mais agradável (*Pleasantness*) do que a emoção "*Kindly*". Como descrevem os autores, emoções de um mesmo octante são altamente correlacionadas positivamente; octantes adjacentes são moderadamente correlacionados; octantes separados por 90 graus não possuem nenhum grau de correlação e octantes separados por 180 graus são opostos e, portanto, possuem uma alta correlação negativa.

A LMMD (*Latin Mood Music Database*), que é a base de músicas utilizada neste trabalho, foi rotulada com base no modelo de Watson e Tellegen (1985). Segundo Santos e Silla (2015), as seis emoções encontradas na base representam seis dos octantes apresentados na estrutura da Figura 2. A emoção "tristeza" representa o octante *Unpleasantness* (desagrádavel). A emoção "decepção" representa o octante *High Negative Affect* (alto afeto negativo). O octante *Strong Engagement* (forte envolvimento) é representado pela emoção "amor". Já a emoção "excitado/entusiasmado" representa o octante *High Positive Affect* (alto afeto positivo). A emoção "paixão" representa o octante *Pleasantness* (agradável) e, por fim, o octante *Low Positive Affect* (baixo afeto negativo) é representado pela emoção "alegria".

2.2 EXTRAÇÃO DE CARACTERÍSTICAS

Diferentes fontes de informações têm sido empregadas para a extração de características visando a construção de classificadores de músicas por emoções. Informações relacionadas aos acordes, notas musicais, metadados e arquivos MIDI das músicas têm sido exploradas, porém as duas fontes de informações mais comumente utilizadas são as letras e os áudios, conforme constatado na revisão sistemática apresentada na Subseção 3.1.4. Também é comum encontrar trabalhos que combinam características extraídas de diferentes fontes de informação. A seguir são apresentados aspectos relacionados à extração de características a partir de letras de música. A extração de características a partir do áudio será apresentada na Subseção 2.2.2.

2.2.1 LETRAS DE MÚSICA COMO FONTE DE INFORMAÇÃO

O uso das letras de música como fonte de informação tem crescido nos últimos anos e inúmeras características têm sido extraídas e utilizadas no processo de classificação. Conforme mostrado na Subseção 3.1, a maioria dos trabalhos utilizam modelos do tipo *bag-of-words* para a extração de características das letras das músicas. Características estilísticas e psicolinguísticas, como contagens de palavras a partir de léxicos de sentimentos, também são frequentemente

utilizadas.

Assim como em outras tarefas de processamento de texto, é comum que as letras sejam pré-processadas antes da extração de características. Duas técnicas comumente utilizadas são a remoção de *stopwords* e o *stemming*. *Stopwords*, também chamadas de palavras funcionais (HU; DOWNIE, 2010a), são palavras que, por si só, não contribuem para a semântica do documento. As *stopwords* dependem do contexto em que estão inseridas e geralmente incluem os artigos e preposições. Mesmo assim, alguns trabalhos, como o de Hu e Downie (2010b), optam por não remover *stopwords* de representações como bigramas e trigramas por essas palavras servirem como ligação entre palavras de conteúdo. *Stemming* se refere a uma versão mais simples do processo de lematização e se resume à extração do radical das palavras a fim de reduzir o vocabulário do texto (JURAFSKY; MARTIN, 1999). Ao reduzir o tamanho do vocabulário, essa etapa de pré-processamento pode auxiliar a diminuir a esparsidade do modelo *bag-of-words*, e consequentemente, reduzir a complexidade computacional.

2.2.1.1 *BAG-OF-WORDS* NORMALIZADOS COM TF-IDF E CARACTERÍSTICAS ESTILÍSTICAS

Na literatura há uma variedade de características que codificam tanto informações superficiais quanto conhecimentos linguísticos mais profundos. Conforme mencionado anteriormente, uma das formas mais comuns de converter letras de músicas em características é pelo uso de modelos *bag-of-words*.

Bag-of-words (BoW) é o conjunto desordenado das palavras de uma coleção de textos utilizado para representação de documentos. Essa forma de representação padroniza os documentos em vetores de palavras que sinalizam a frequência de ocorrência dessas palavras no documento. Assim, são ignoradas as posições das palavras mantendo apenas as suas frequências no texto (JURAFSKY; MARTIN, 1999).

Os chamados n-gramas são formas de *bag-of-words* e podem ser definidos como uma sequência de *n* palavras ou caracteres (JURAFSKY; MARTIN, 1999). São encontrados na literatura o uso predominante de 1-gramas (unigramas), 2-gramas (bigramas), 3-gramas (trigramas) ou ainda 4-gramas (quadrigramas). No Quadro 2 são apresentados dois exemplos nos quais as sentenças são divididas em bigramas. Nas duas últimas linhas do quadro são descritos os vetores de bigramas gerados a partir das sentenças 1 e 2, respectivamente.

Quadro 2: Exemplos de bigramas.

Sentença 1: Eu ainda vou poder cantar
Sentença 2: Eu ainda vou poder lembrar
Bigramas da sentença 1 (S1) (‘Eu ainda’, ‘ainda vou’, ‘vou poder’, ‘poder cantar’)
Bigramas da sentença 2 (S2) (‘Eu ainda’, ‘ainda vou’, ‘vou poder’, ‘poder lembrar’)

Fonte: Autoria Própria (2019)

Após a geração dos bigramas, as suas respectivas frequências são calculadas. O resultado para o exemplo do Quadro 2 é apresentado na Tabela 1.

Tabela 1: Representação por bigramas das sentenças da Figura 2.

Sentença	Eu ainda	ainda vou	vou poder	poder cantar	poder lembrar
S1	1	1	1	1	0
S2	1	1	1	0	1

Fonte: Autoria Própria (2018)

A geração de n-gramas também pode ser feita por caracteres em vez de palavras. Segundo Li et al. (2015), letras de músicas não são regulares e, portanto, podem conter erros de segmentação quando a mesma é realizada por palavras. A segmentação por caracteres, segundo os autores, corrigiu esses erros em letras de músicas em chinês. Essa também foi a abordagem utilizada por Ribeiro (2015) e Przybysz (2016), que estudaram letras de músicas latinas.

O modelo *bag-of-words* baseado na frequência de n-gramas pode não ser muito discriminativo. Se um determinado n-grama aparece com uma alta frequência em vários dos documentos que precisam ser classificados, esse n-grama não auxiliará o processo de classificação. Em contra partida, n-gramas com alta frequência são mais relevantes do que n-gramas que ocorrem poucas vezes (JURAFSKY; MARTIN, 1999). Para balancear esses dois aspectos, utiliza-se medida TF-IDF, que corresponde ao produto de dois valores, a saber: a frequência dos termos (TF) e a inverso da frequência dos documentos (IDF).

A TF corresponde à frequência dos n-gramas no documento. Normalmente essa frequência é diminuída pelo fato de que um n-grama que apareça 100 vezes em um documento não faz esse n-grama ser 100 vezes mais provável ou útil para o significado do

documento (JURAFSKY; MARTIN, 1999). Uma forma de diminuir o valor da frequência proporcionalmente é normaliza-la. Assim, a TF pode ser definida como:

$$TF = \frac{freq(t,d)}{maxOthers(t,d)} \tag{1}$$

em que $freq(t,d)$ é a frequência do n-grama t no documento d e $maxOthers(t,d)$ é a maior frequência de um n-grama no documento d.

A IDF é baseada na frequência por documentos que um n-grama ocorre. Esse valor é usado para atribuir um peso maior para n-gramas que ocorrem em poucos documentos da coleção (JURAFSKY; MARTIN, 1999). Assim, a IDF pode ser definida como

$$IDF(t) = log\frac{N}{df(t)} \tag{2}$$

em que N é o total de documentos da coleção e $df(t)$ é o número documentos em que o n-grama t ocorre.

Por meio da combinação dos valores de TF e IDF, tem-se a fórmula da medida TF-IDF:

$$\text{TF-IDF}(t,d) = \text{TF})(t,d) * \text{IDF}(t) \tag{3}$$

Após o cálculo do TF-IDF é comum normalizar os vetores pela norma euclidiana transformando-os assim, em vetores unitários. A norma euclidiana pode ser definida como:

$$v_{norm} = \frac{v}{\sqrt{v_1^2 + v_2^2 + ... + v_n^2}} \tag{4}$$

em que v são os valores do vetor.

Outra categoria comum de características são as chamadas de estilísticas. Elas capturam estatísticas do texto, como o número de palavras, o número de caracteres especiais, o número total de linhas, o número total de palavras únicas, entre outras (HU; DOWNIE, 2010b). Por serem características relacionadas diretamente à escrita, independentes de língua ou de qualquer informação semântica, são características relativamente simples de serem extraídas. Conforme apresentado na Subseção 3.1, seis dos trabalhos encontrados na literatura utilizaram esse tipo de característica, além dos trabalhos de Ribeiro (2015) e Przybysz (2016).

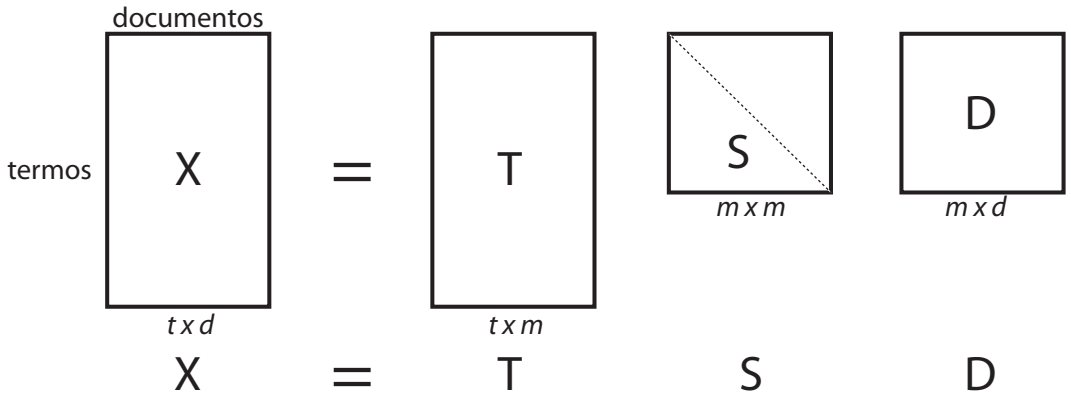
2.2.1.2 ANÁLISE DE SEMÂNTICA LATENTE

A Análise de Semântica Latente (LSA) é uma técnica estatística desenvolvida por Deerwester et al. (1990) para extrair e inferir relações do uso contextual de palavras. Essa técnica foi criada no contexto de recuperação de informação e sua ideia principal consiste em formar um espaço semântico em que a relação entre as palavras se dá pela ocorrência em contextos comuns (SOUZA, 2011). Além disso, por meio dessa técnica é possível reduzir a dimensionalidade de uma matriz gerada por modelos de BoW. Por exemplo, em um determinado contexto as palavras "paladino" e "herói" podem ser consideradas sinônimas. Se gerada uma matriz BoW com esse contexto, essas palavras ocupariam duas colunas. Com a aplicação da LSA, essas duas palavras são representadas em uma única coluna, diminuindo assim a dimensionalidade e, consequentemente, a computação necessária para processar a matriz (ITO, 2018).

A aplicação da técnica consiste em gerar uma matriz TF-IDF e então utilizar a Decomposição de Valores Singulares (SVD), técnica derivada da álgebra linear, para reduzir a dimensionalidade da matriz e encontrar os principais padrões associativos (SOUZA, 2011). A seguir é descrito o processo de como a SVD decompõe uma matriz TF-IDF conforme apresentado em SOUZA (2011).

Considere uma matriz $X_{t,d}$ normalizada com TF-IDF, já **transposta**, em que t é o número de termos (linhas) e d é o número de documentos (colunas). A matriz X é decomposta em outras três matrizes de forma que, $X = TSD$. A Figura 3 apresenta o esquema da decomposição da matriz X.

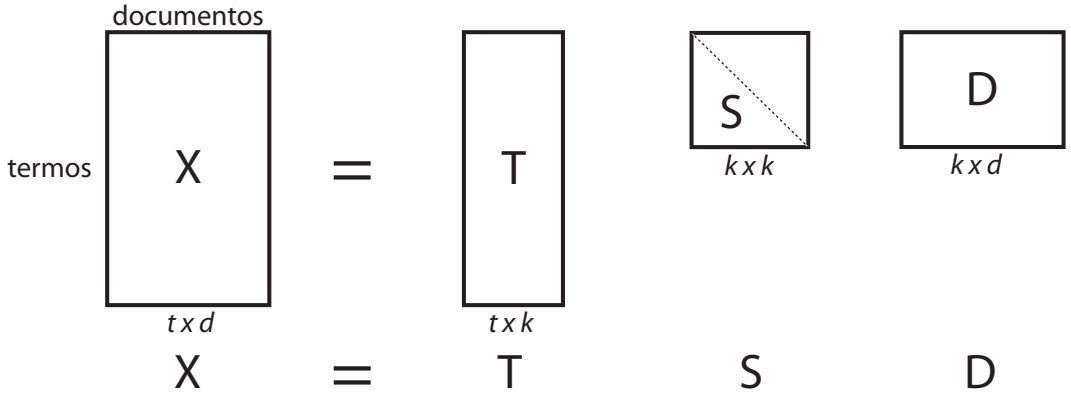
Figura 3: Decomposição da matriz X em três matrizes.



Fonte: Deerwester et al. (1990)

Na Figura 3, m é o número de dimensões, sendo $m \leq min(t,d)$. $T_{t,m}$ é a matriz de vetores singulares à esquerda, $S_{m,m}$ é a diagonal principal de valores singulares em ordem decrescente e $D_{m,d}$ é matriz de vetores singulares à direita. Após essa decomposição, a matriz S é reduzida substituindo m por um valor k em que, $k < m$. Os $m-k$ menores valores da diagonal S são removidos, pois são os menos relevantes, uma vez que a matriz estava em ordem decrescente conforme sua diagonal principal. Consequentemente, as linhas e colunas correspondentes nas matrizes T e D também são removidas. Esse processo de redução é ilustrado na Figura 4. Dessa forma, a LSA reduz a matriz X em seus termos mais significativos e reduz o tempo computacional de seu processamento.

Figura 4: Decomposição da matriz X em três matrizes.



Fonte: Deerwester et al. (1990)

2.2.1.3 *WORD VECTORS* E *PARAGRAPH VECTORS*

Os n-gramas têm sido muito utilizados na literatura por causa da sua facilidade de aplicação e por vezes sendo muito eficientes quanto ao desempenho na classificação. Contudo, essa forma de representação de textos possui duas limitações principais: a ordem das palavras é perdida e a semântica das palavras é ignorada (LE; MIKOLOV, 2014). Nesse contexto, Mikolov et al. (2013) propuseram uma nova forma de representar palavras em um espaço vetorial, chamada *word vectors* (Word2Vec).

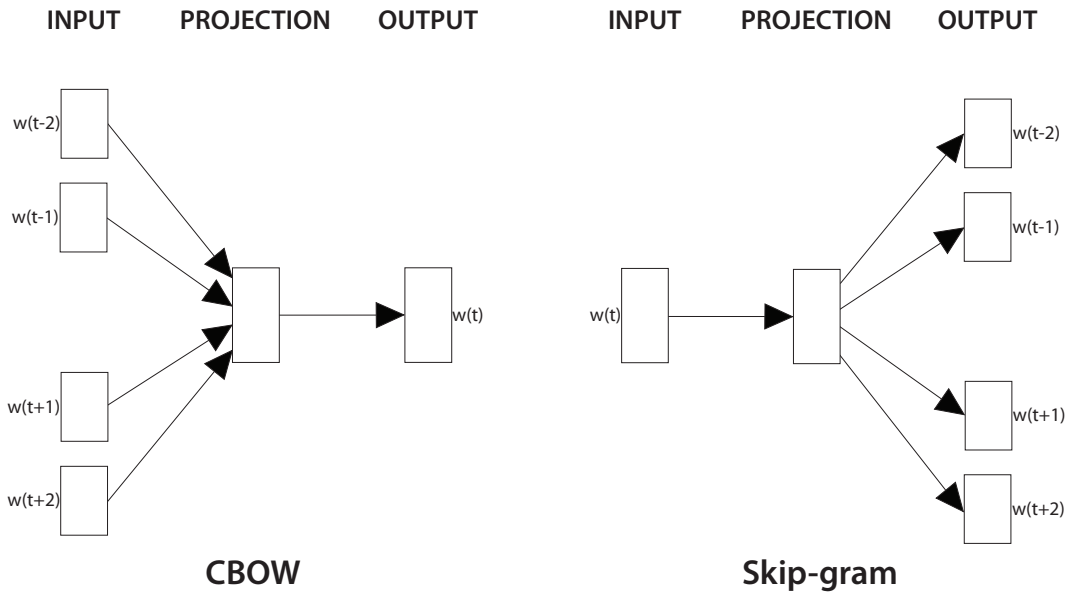
Segundo os autores, Word2Vec foca em representações distribuídas das palavras aprendidas por uma rede neural treinada usando gradiente descendente estocástico e *backpropagation*. Por meio dessa rede, as palavras são distribuídas em um espaço vetorial de maneira que palavras com semântica similar tendem a ficar mais próximas uma das outras. Por exemplo, levando em consideração as palavras *powerful*, *strong* e *Paris*, as duas primeiras

tendem a ficar mais próximas no espaço vetorial do que a primeira e a terceira palavra. Isso acontece porque as palavras *powerful* e *strong* têm semântica similar.

O modelo proposto por Mikolov et al. (2013) não contém a camada oculta da rede neural *feedforward* reduzindo assim a complexidade computacional da rede. Embora isso não permita que a rede represente dados tão precisamente quanto as redes *feedforward*, em contra partida elas podem ser treinadas com mais dados de forma eficiente. Dessa maneira, a rede contém as camadas de entrada, projeção e saída, sendo a camada de projeção uma matriz de pesos e, não composta por neurônios como as demais redes neurais. Cada coluna dessa matriz representa uma palavra do vocabulário e cada linha representa uma dimensão na qual se quer projetar a palavra. A quantidade de dimensões para representar palavras pode ser escolhida arbitrariamente.

Nesse sentido, dois modelos foram propostos, a saber: *Continuous bag-of-words* (CBOW) e Skip-gram. O primeiro utiliza um contexto de *N* palavras para predizer uma palavra e o último utiliza uma única palavra para predizer as *N* palavras do contexto.

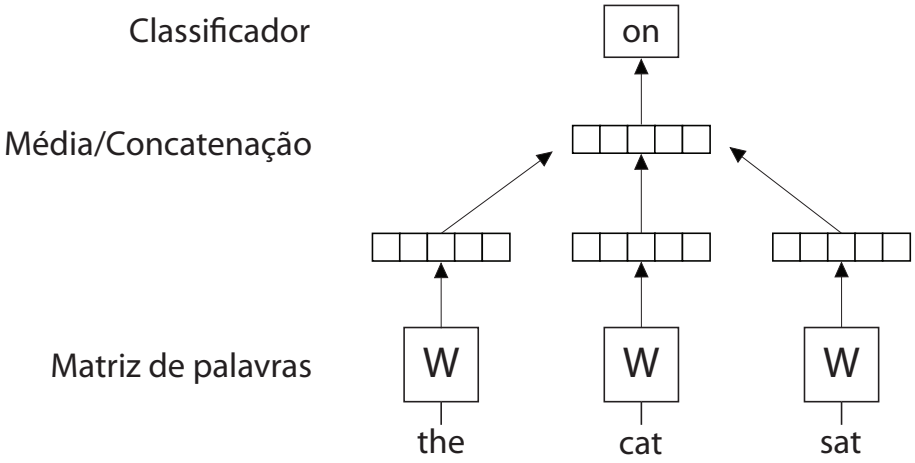
Figura 5: Arquiteturas do CBOW e Skip-gram.



Fonte: Mikolov et al. (2013)

Como pode ser observado na Figura 5, no CBOW, duas palavras anteriores, $w(t-2)$ e $w(t-1)$, e duas palavras posteriores, $w(t+1)$ e $w(t+2)$, são utilizadas para predizer a palavra central, $w(t)$. No modelo Skip-gram o processo é invertido, a palavra $w(t)$ é utilizada para predizer as demais. Um outro exemplo apresentado em Le e Mikolov (2014) é exibido na Figura 6.

Figura 6: *Framework* do *Word Vectors*.

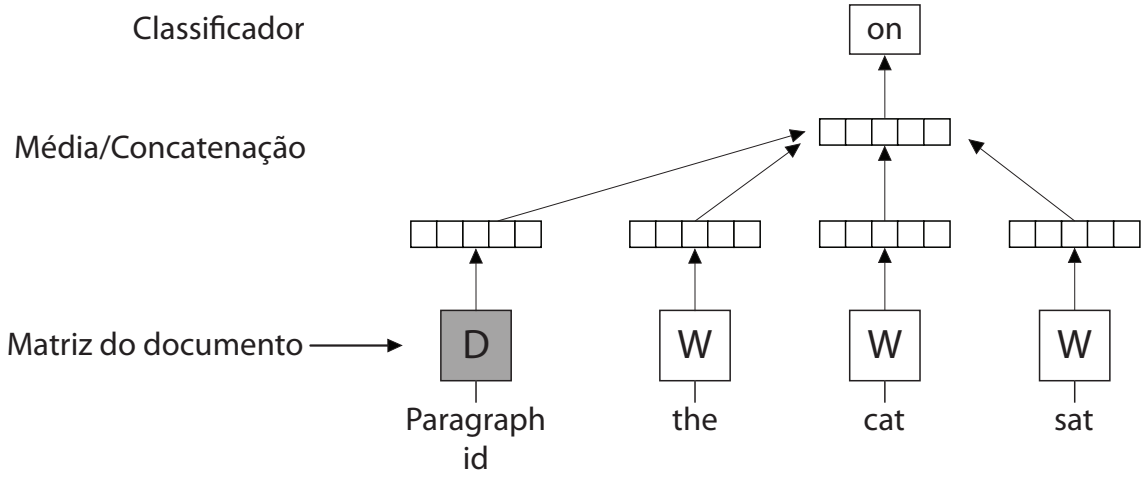


Fonte: Adaptado de Le e Mikolov (2014)

O *framework* apresentado na Figura 6 leva em consideração três palavras anteriores (*the*, *cat*, *sat*) para predizer a quarta palavra, *on*. Nesse *framework*, cada palavra é mapeada para um vetor exclusivo, representado por uma coluna em uma matriz W (matriz cuja as palavras são representadas). No início do aprendizado, esses vetores podem ser gerados de diversas formas, por exemplo de maneira aleatória. Para predizer a quarta palavra, todos os vetores de W são mapeados para um mesmo vetor por meio de uma concatenação ou soma. O vetor resultante é então comparado com o vetor da palavra a ser predita. Depois disso, os erros são propagados pela rede e os pesos da camada de projeção são corrigidos assim como os valores da matriz W.

Inspirados nesse modelo, Le e Mikolov (2014) propuseram o *Paragraph Vector*. O termo parágrafo é utilizado para representar qualquer segmento de um texto, podendo ser uma sentença, um parágrafo ou um documento. Esse método segue a mesma ideia do *Word Vectors*, mapeando cada palavra para um vetor exclusivo e, além disso, mapeando também cada parágrafo para um vetor. A seguir, tanto os vetores das palavras quanto o vetor do parágrafo são concatenados para definir o vetor da palavra que se está querendo predizer. Na Figura 7 é apresentado o *framework* proposto por Le e Mikolov (2014).

Figura 7: *Framework* do *Paragraph Vectors*.



Fonte: Adaptado de Le e Mikolov (2014)

Comparando-se as Figura 7 e Figura 6 é possível perceber a diferença entre os modelos. Conforme dito anteriormente, no *Paragraph Vector*, além dos vetores das palavras, o próprio vetor do parágrafo é utilizado na concatenação para predizer a palavra. Os parágrafos são mapeados em uma matriz D, assim como as palavras são mapeadas em uma matriz W. Como os autores ressaltam, o vetor do parágrafo pode ser visto como uma memória que serve para armazenar o contexto atual das palavras. Por essa razão, esse modelo pode ser chamado de Modelo de Memória Distribuída do Vetor de Parágrafo (*Distributed Memory Model of Paragraph Vectors* - PV-DM). É importante ressaltar que o *Paragraph Vector* também possui as duas arquiteturas CBOW e Skip-Gram.

2.2.2 ÁUDIOS DE MÚSICAS COMO FONTE DE INFORMAÇÃO

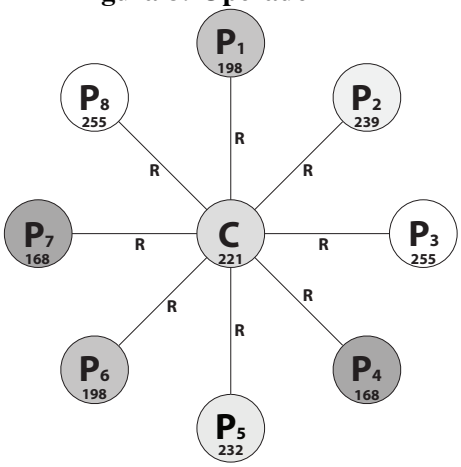
Além das letras, o áudio também é uma fonte de informação comumente utilizada no contexto de recuperação de informação de música. É possível extrair características diretamente do sinal de áudio ou ainda transformar o sinal de áudio em espectrogramas para posterior extração de características. A principal diferença dessa transformação é proporcionar a extração de características baseadas em atributos visuais da imagem, como a textura (TAVARES et al., 2017). Pode-se definir espectrograma como uma representação visual do espectro de frequência do som e que tem como principal atributo a textura (COSTA, 2013). No contexto de classificação de áudio, é comum que as características extraídas dos espectrogramas sejam chamadas de visuais e as características extraídas diretamente do sinal de áudio sejam chamadas de acústicas.

Existem diversos descritores para extrair características visuais e acústicas. Neste trabalho, o *Robust Local Binary Pattern* (RLBP), que é descrito na Subseção 2.2.2.1, foi utilizado para extrair características visuais. Já as características acústicas foram exploradas utilizando as Mel-frequency cepstral coefficients (MFCC) combinadas com *Rolloff*, *Spectral centroid*, *Flux* e *Zerocrossings* e os descritores *Rhythm Patterns* (RP), *Statistical Spectrum Descriptor* (SSD) e *Rhythm Histogram* (RH). Todas essas características são apresentadas na Subseção 2.2.2.2.

2.2.2.1 *LOCAL BINARY PATTERN* (LBP)

O *Local Binary Pattern* (LBP) é um método eficiente para extrair características relacionadas a textura de imagens digitais, e por isso é apropriado para descrever conteúdo de espectrogramas. O método tem como resultado um histograma de padrões binários locais que servem como descritores da textura. Nesse método, para todo *pixel* central C e seus *pixels* vizinhos P equidistantes a uma distância R, os padrões são extraídos a partir da diferença de intensidade entre C e seus vizinhos P. A quantidade de padrões que podem ser extraídos depende do valor atribuído a P. Por exemplo, se $P = 8$ existem 2^8 de padrões possíveis (TAVARES et al., 2017). A Figura 8 mostra o exemplo descrito do operador LBP.

Figura 8: Operador LBP



Fonte: Adaptado de Costa (2013)

Como descrito no parágrafo anterior, os padrões são extraídos de acordo com a diferença de intensidade dos *pixels* P para o *pixel* C. Assim, para cada *pixel* em P é calculada a diferença de intensidade em relação ao *pixel* central C atribuindo o valor 1 se a intensidade do *pixel* central for maior ou igual a do *pixel* Pi e o valor 0, caso contrário, como descrito na

Equação 5

$$s(g_i - g_C) \begin{cases} 1 & \text{se} \quad g_C \geq g_i, \\ 0 & \text{se} \quad g_C < g_i \end{cases} \tag{5}$$

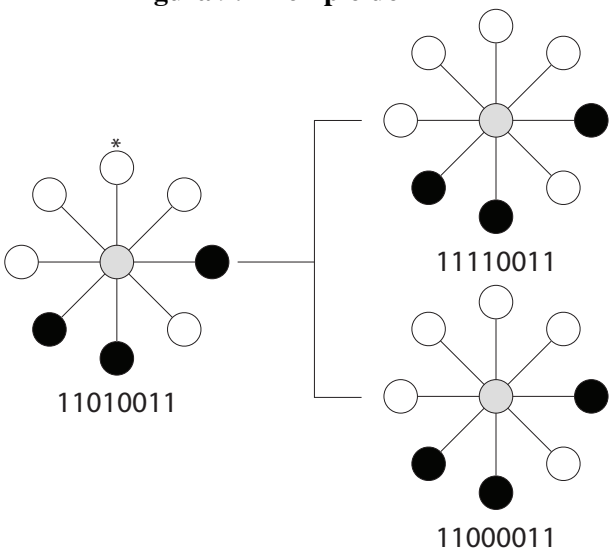
em que, g_C é a intensidade do nível de cinza do *pixel* central C e g_i corresponde a intensidade de nível de cinza dos vizinhos P.

Analisando o exemplo da Figura 8, podemos extrair o padrão binário 10010110 assumindo que o *pixel* $P1$ é comparado por primeiro. É válido lembrar que o operador LBP segue o sentido horário.

A partir disso, Ojala et al. (2002) introduziram o conceito de uniformidade dos padrões que diz que um padrão é uniforme se o número de transações entre 1 e 0 no padrão for menor ou igual a 2, considerando o padrão como uma lista circular. Dessa forma, não é necessário utilizar todos os padrões do histograma encontrado, mas somente os padrões uniformes, gerando assim, um vetor de menor dimensionalidade. Pode-se dar como exemplos de padrão uniformes, 00010000, 11111111 e 11110000. A versão mais tradicional do LBP apresentado na Figura 8 contém 58 possíveis padrões uniformes e, além disso, todos os outros padrões não uniformes encontrados são contabilizados em uma última coluna do histograma.

Segundo Chen et al. (2013), o LBP pode ser sensível à ocorrência de alguns ruídos na imagem. Por causa disso, os autores propuseram o *Robust Local Binary Pattern* (RLBP), uma variação do LBP, que trata a questão da uniformidade de uma forma diferente. Esse método considera padrões não uniformes que podem ser convertidos em padrões uniformes com a mudança de apenas um dígito. Na Figura 9 é apresentado um exemplo.

Figura 9: Exemplo do RLBP



Fonte: Adaptado de Costa et al. (2017)

O *pixel* demarcado com "*" foi considerado com o *pixel* inicial. Como pode ser observado, o padrão 11010011 é considerado não uniforme por causa do terceiro ou quarto *pixel*. Com a alteração de um desses *pixels* o padrão se torna uniforme podendo ser 11110011 (alteração do terceiro *pixel*) ou 11000011 (alteração do quarto *pixel*). Os experimentos apresentados pelos autores comprovaram que o RLBP superou o LBP e outros descritores de textura, principalmente quando ruídos eram adicionados às imagens.

2.2.2.2 CARACTERÍSTICAS ACÚSTICAS

Nesta subseção são apresentadas as características que foram extraídas diretamente dos áudios das músicas. As MFCCs, *Rolloff*, *Spectral centroid*, *Flux* and *Zerocrossings* compõem um conjunto de características, enquanto que, os descritores RP, RH e SSD compõem outro conjunto de características acústicas.

O MFCC é um descritor acústico proposto por Mermelstein (1976) utilizado para reconhecimento de áudio. Esse descritor se baseia na forma espectral das ondas de áudio para extrair o vetor de características. O processo de extração inicia com o sinal do áudio sendo dividido em *frames* com intervalos de 20 a 40 milisegundos, geralmente. A partir dessa divisão, para cada *frame* gerado é calculado uma Transformada de Fourier Discreta (DFT) a fim de obter a representação do áudio no domínio da frequência. Depois disso, operações logarítmicas são aplicadas para extrair a amplitude de cada *frame*. Por fim, a frequência Mel é obtida por meio de uma função *smooting* e os coeficientes são calculados por meio de uma Transformada Discreta de Cosseno (DCT). Esses coeficientes compõem o vetor de características (PAULINO et al., 2018).

Na literatura é possível encontrar diversos trabalhos que combinam algumas características com as MFCCs (TZANETAKIS; COOK, 2002). Algumas dessas características são: *Rolloff*, *Spectral centroid*, *Flux* and *Zerocrossings*. Por isso, essas características também foram utilizadas neste trabalho e estão descritas em Tzanetakis e Cook (2002).

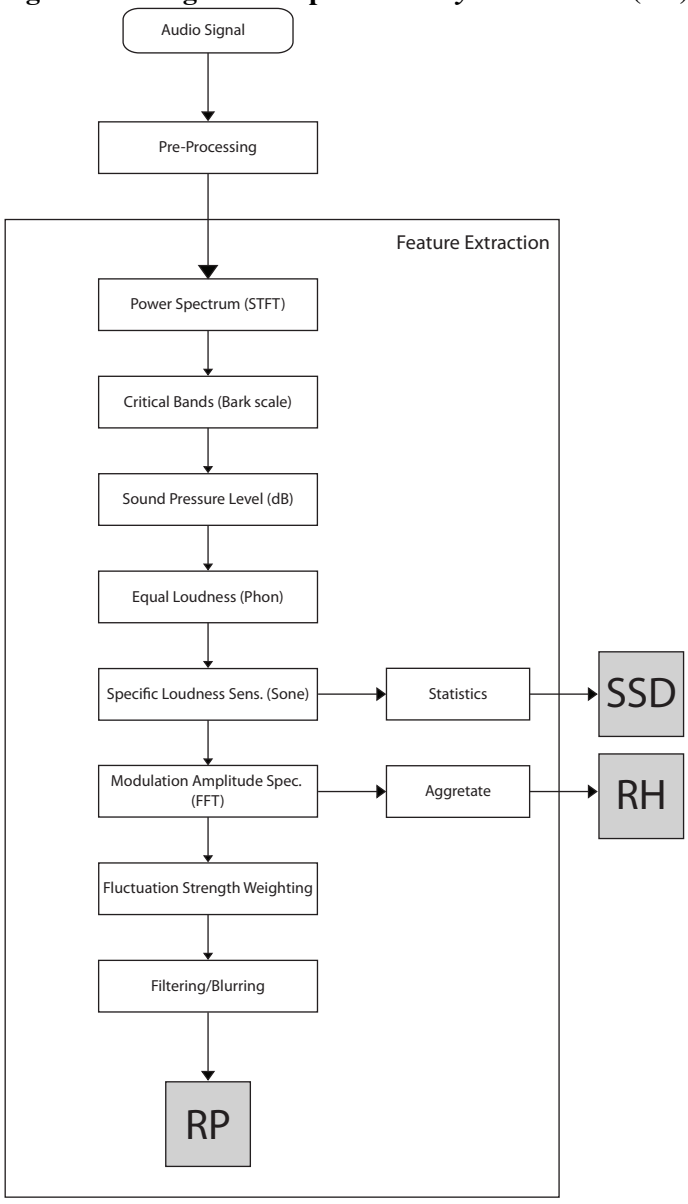
Rhythm Patterns (RP) é um processo que descreve a amplitude da modulação para uma faixa de frequências de modulação em bandas críticas (LIDY; RAUBER, 2005) e, a partir disso, é possível realizar a extração de características acústicas. Na Figura 10 é ilustrado o processo RP.

Antes da extração das características, os áudios são preprocessados. Esse preprocessamento consiste em converter os áudios para um dos formatos wav, mp3 ou au, realizar a média de múltiplos canais (se houver múltiplos) do áudio para apenas um canal e

extrair uma faixa de 6 segundos do áudio. Após esse preprocessamento, o segmento do áudio é transformado usando uma *Fast Fourier Transform* (FFT) e, a partir disso, a escala de Bark é aplicada para agrupar as bandas de frequência em 24 bandas críticas. A partir dessas bandas são extraídas as seguintes medidas estatísticas: média, mediana, variância, assimetria, curtose, valor mínimo e máximo. Os valores de energia nas bandas são transformados para decibéis [dB], os níveis de sonoridade são calculados e a sensação específica de sonoridade por banda crítica é computada. Uma nova FFT é aplicada ao sonograma da escala de Bark para obter uma invariante de tempo. Por fim, um filtro de gradiente e suavização gaussiana são então aplicados para melhorar a semelhança entre os padrões de ritmo e, então, as medianas dos padrões de ritmo por segmento são calculadas (COSTA et al., 2017).

De acordo com Lidy e Rauber (2005), medidas estatísticas são capazes de descrever o conteúdo do áudio por meio da ocorrência de batidas ou outra variação rítmica de energia a partir das 24 bandas críticas de Bark. Essas medidas são utilizadas pelo descritor *Statistical Spectrum Descriptor* (SSD) que pode ser entendido como uma primeira parte do processo RP. Com esse descritor são extraídas 168 características considerando o ritmo e timbre do áudio. Um segundo descritor interno do processo RP, denominado *Rhythm Histogram* (RH) também é proposto pelos autores. Ao contrário do RP e SSD, o RH soma as magnitudes de cada compartimento de frequência de modulação de todas as 24 bandas críticas em vez de armazenar as informações por banda crítica. Como apresentado em Lidy e Rauber (2005), é gerado um vetor de 60 dimensões a partir desse descritor.

Figura 10: Diagrama do processo *Rhythm Patterns* (RP)



Fonte: Adaptado de Lidy e Rauber (2005)

2.3 DESBALANCEAMENTO DE BASES DADOS E O ALGORITMO SMOTE

Em Aprendizagem de Máquina (AM) é comum utilizar bases de dados para treinar e testar diversos tipos de modelos de aprendizagem. Essas bases, por vezes, são desbalanceadas, o que pode dificultar o processo de aprendizagem. Considera-se que uma base de dados é desbalanceada quando as classes presentes na base não estão igualmente (ou aproximadamente) representadas (CHAWLA et al., 2002). Para atenuar esse problema, algoritmos como o *Synthetic Minority Over-sampling Technique* (SMOTE) (CHAWLA et al., 2002), implementam técnicas que permitem balancear, ainda que artificialmente, a base de dados.

O SMOTE tem como ideia principal criar exemplos sintéticos das classes minoritárias a partir dos exemplos já encontrados na base de dados. O algoritmo tem como entrada um conjunto T composto pelas amostras da classe minoritária, uma quantidade N de amostras sintéticas que se deseja criar e uma quantidade k de vizinhos mais próximos. A seguir são apresentados os passos do algoritmo de forma simplificada e um exemplo de sua aplicação. O algoritmo completo pode ser encontrado em Chawla et al. (2002).

1. Para cada amostra i pertencente a T:
   (a) Armazene os k vizinhos mais próximos de i em um vetor $i.nnarray$.
   (b) Escolha aleatoriamente um vizinho n em $i.nnarray$.
   (c) Considere s como a amostra sintética a ser gerada.
   (d) Para cada atributo att em i:
      i. Calcule a diferença d_{ni}, sendo $d_{ni} = n.att - i.att$.
      ii. Escolha aleatoriamente um número r entre 0 e 1.
      iii. O atributo da amostra sintética é $s.att = i.att + r * d_{ni}$.

Os passos 1b e seguintes são executados N vezes já que essa é a quantidade de amostras sintéticas a serem geradas. Para facilitar o entendimento do algoritmo é apresentado a seguir um exemplo.

Considere a como uma amostra de uma classe minoritária. A amostra a contém os atributos $x = 1$ e $y = 1$. Considere também $k = 1$, $N = 1$ e $r = 0,5$ (para o cálculos do dois atributos). Por fim, considere que b é o vizinho mais próximo de a e que seus atributos são $x = 2$ e $y = 2$. Assim, $a.nnarray = [b]$. A partir do passo 1b, seguem os cálculos para os atributos x e y, respectivamente. A Equação 6 apresenta o cálculo da distância entre a e b levando em consideração apenas o atributo x. Na Equação 7 é calculado o valor do atributo x para a amostra sintética s. De maneira análoga, o mesmo o ocorre nas Equações 8 e 9 para o atributo y da amostra sintética s.

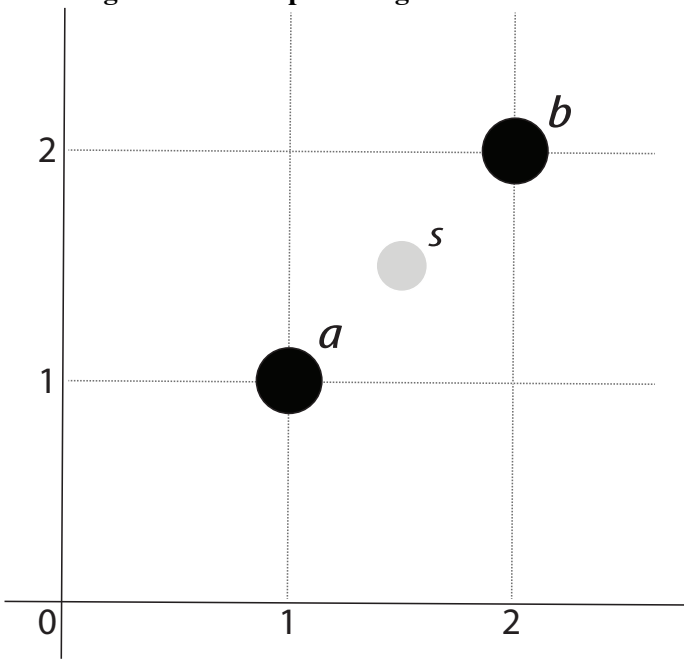
$$\begin{aligned} d_{ba} &= b.x - a.x \\ d_{ba} &= 2 - 1 \\ d_{ba} &= 1 \end{aligned} \qquad (6) \qquad \begin{aligned} s.x &= a.x + r * d_{ba} \\ s.x &= 1 + 0,5 * 1 \\ s.x &= 1,5 \end{aligned} \qquad (7)$$

$$d_{ba} = b.y - a.y \qquad (8)$$
$$d_{ba} = 2 - 1$$
$$d_{ba} = 1$$

$$s.y = a.y + r * d_{ba} \qquad (9)$$
$$s.y = 1 + 0,5 * 1$$
$$s.y = 1,5$$

Portanto, $s.x = s.y = 1,5$. Na Figura 11, é possível visualizar o exemplo apresentado.

Figura 11: Exemplo do algoritmo SMOTE.



Fonte: Autoria Própria (2020)

2.4 ALGORITMOS DE CLASSIFICAÇÃO

Problemas de classificação podem ser encontrados em diversos ramos da sociedade. Como apresenta Abu-Mostafa et al. (2012), esses problemas variam de previsões financeiras e diagnósticos médicos até métodos para classificar filmes ou músicas e recomenda-las a usuários. Nesses problemas, as amostras (filmes, músicas, etc.) são coletadas, observadas e, a partir disso, é possível atribuir classes a essas amostras. Por exemplo, no problema de classificação de filmes por gêneros, um determinado filme (amostra) pode ser classificado em suspense ou comédia (classes).

O problema de classificação em muitos casos pode se tornar algo complexo, principalmente, quando existem muitas amostras a serem analisadas. Levando em consideração

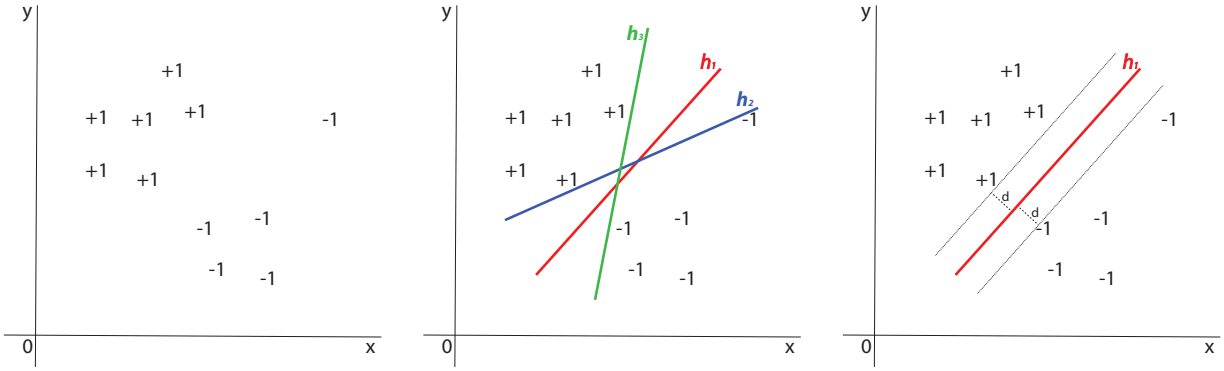
o exemplo do parágrafo anterior, a tarefa de classificar alguns filmes por gênero pode ser trivial, mas à medida que a quantidade de filmes aumenta o problema vai se tornando inviável de ser realizado manualmente. Nesse sentido, a área de Aprendizagem de Máquina (AM) tem como objetivo automatizar esse processo de classificação e, consequentemente, viabilizar a classificação de milhares de amostras em curto prazo de tempo. Para isso, um conjunto de amostras é observado e de cada amostra são coletados conjuntos de valores de atributos que são usados como entrada em algoritmos de aprendizagem (também chamados de indutores). *Support Vector Machine* (SVM) e algoritmos baseados em árvores de decisões são exemplos desses algoritmos. Esses algoritmos de aprendizagem, geralmente, tem como saída classificadores (também chamados de modelos) e esses por sua vez, são capazes de classificar amostras não conhecidas previamente de acordo com as amostras observadas no algoritmo de aprendizagem.

Nesta subseção são apresentados os algoritmos que foram utilizados no desenvolvimento deste trabalho. *Support Vector Machine* (SVM) foi o principal classificador utilizado e é apresentado na Subseção 2.4.1. Foram explorados também os algoritmos *ensemble Adaboost* e *Bagging*. Os algoritmos *Decision Stump*, *Random Forest*, *Random Tree* e *REPTree* foram utilizados como algoritmos base para os *ensemble*. Todos esses algoritmos são apresentados na Subseção 2.4.2.

2.4.1 *SUPPORT VECTOR MACHINE* (SVM)

Proposto por Vapnik (1995), o *Support Vector Machine* (SVM) é um dos algoritmos mais conhecidos de AM e tem sido aplicado em diversos domínios para problemas de classificação. Isso porque apresenta vantagens, como pouca complexidade no ajuste de parâmetros e uso adequado em espaços de muitas dimensões (DHANARAJ; LOGAN, 2005). O SVM tem como princípio criar um hiperplano ótimo que possa separar os exemplos de classes em um espaço n-dimensional. Por criar hiperplanos o SVM é considerado um classificador linear. Pode-se entender o hiperplano ótimo como aquele que é equidistante entre os exemplos das classes. Considere como exemplo um conjunto de amostras que está separado em duas classes $\{+1, -1\}$. Dispondo essas amostras em um plano bidimensional, pode-se obter o resultado apresentado na Figura 12a. É necessário ressaltar que, como apresenta o exemplo dessa figura, as classes são linearmente separáveis. Como é apresentado na Figura 12b, vários hiperplanos podem ser propostos para separar as amostras e todos possuem margens iguais entre as amostras das classes. Porém, o algoritmo do SVM visa maximizar essas margens para encontrar o hiperplano ótimo. Por causa disso, o hiperplano h_1 foi escolhido.

Figura 12: Hiperplanos do SVM.

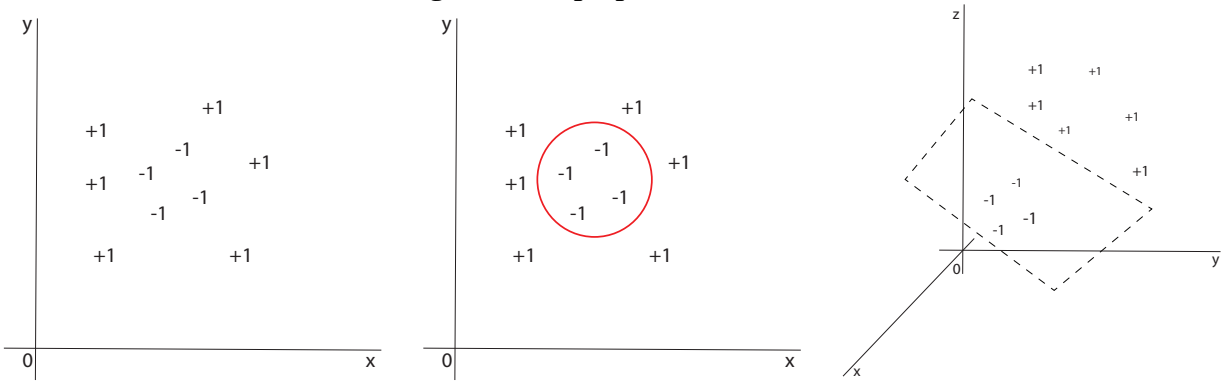


(a) Disposição das amostras. (b) *n* hiperplanos. (c) Hiperplano ótimo com maximização das margens.

Fonte: Autoria Própria (2020)

Em muitos casos, os dados de treinamento estão dispostos de tal forma que não é possível separá-los linearmente. O SVM também trata esses problemas. Para isso, os dados de treinamento que estão no espaço de entrada são dispostos em um novo espaço de maior dimensão, denominado espaço de características. Essa transformação visa dispor os dados de tal forma que eles possam ser separados por um SVM. A Figura 13 ilustra essa transformação.

Figura 13: Hiperplanos do SVM.



(a) Conjuntos de dados não linear. (b) Fronteira não linear no espaço de entradas. (c) Fronteira linear no espaço de características.

Fonte: Adaptado de Lorena e Carvalho (2007)

A formalização matemática do modelo e aprendizado do SVM pode ser encontrada em Lorena e Carvalho (2007). Cabe destacar que, ainda que com um custo computacional alto, o SVM tornou-se um algoritmo popular, uma vez que apresenta bom desempenho nos mais diversos domínios em que é aplicado. Conforme será apresentado na Subseção 3.1, o SVM foi o algoritmo mais utilizado pelos trabalhos incluídos na revisão sistemática realizada neste trabalho.

2.4.2 *ENSEMBLE* E ALGORITMOS BASE

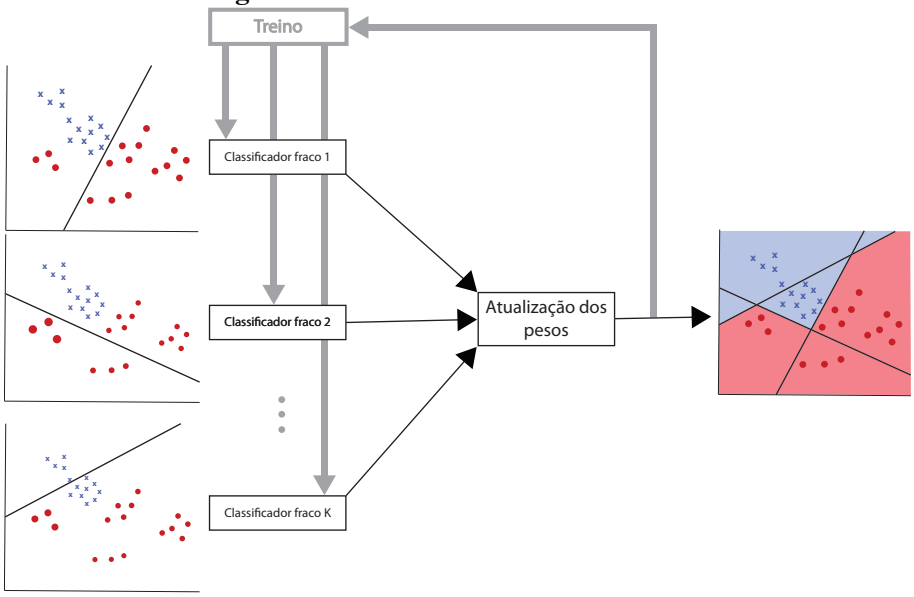
Ensemble é um paradigma de aprendizado de máquina no qual vários modelos são treinados para resolver um mesmo problema. Em contraste com modelos comuns de aprendizado que testam uma única hipótese a partir dos dados de treino, os *ensemble* constroem um conjunto de hipóteses e as combinam para classificar as amostras (ZHOU, 2009). Geralmente, esse conjunto de hipóteses tende à diversidade garantindo que os modelos gerados tenham erros diferentes na classificação e, assim, possam ser complementares entre si.

O algoritmo empregado para construir as hipóteses utilizadas pelo *ensemble* é chamado de algoritmo base que, geralmente, são algoritmos com pouca complexidade para gerar uma variedade maior de modelos e, assim, garantir a diversidade de classificadores. *Decision Stump* é um exemplo desses algoritmos. Ainda assim, algoritmos mais complexos podem ser usados como algoritmos base. É o caso de Ujlambkar e Attar (2012), que utilizou o *ensemble Bagging* com *Random Forest* atingindo os melhores resultados do trabalho. A seguir, são apresentados dois dos principais *ensemble* encontrados na literatura, o *AdaBoost*, algoritmo baseado em *boosting*, e o *Bagging*. Também são descritas as principais características de alguns algoritmos base que são utilizados nesses *ensemble*.

A ideia principal de algoritmos *boosting* é combinar classificadores fracos e imprecisos em um único modelo cujas as predições gerais sejam precisas (SCHAPIRE; FREUND, 2012). Nessa técnica, os classificadores fracos são gerados sequencialmente de forma que um classificador é influenciado pelos resultados de seus antecessores. Um dos algoritmos mais conhecidos baseado em *boosting* é o *AdaBoost* (acrônimo de ***Ada****ptive* ***Boost****ing*). Proposto por Freund e Schapire (1995), o *Adaboost* recebe esse nome porque, diferente dos outros algoritmos de *boosting*, se adapta aos erros das hipóteses retornadas pelos classificadores base. Essa adaptação é realizada atribuindo pesos às amostras da base e alterando esses pesos conforme o aprendizado.

A princípio, todas as amostras recebem o mesmo peso e são, então, utilizadas para induzir um classificador. Amostras corretamente classificadas têm o seu peso diminuído, enquanto amostras incorretamente classificadas têm seu peso aumentado. Na iteração seguinte, um novo classificador é induzido. Esse classificador leva em consideração as amostras com maior peso e, portanto, tenta classifica-las corretamente. Novamente, após a classificação, os pesos das amostras são atualizados de acordo com os erros e acertos e, na iteração seguinte, um novo classificador é modelado. Esse processo é ilustrado na Figura 14 e se repetirá por uma quantidade determinada de iterações.

Figura 14: Funcionamento do *Adaboost*.



Fonte: Adaptado de Wang et al. (2015)

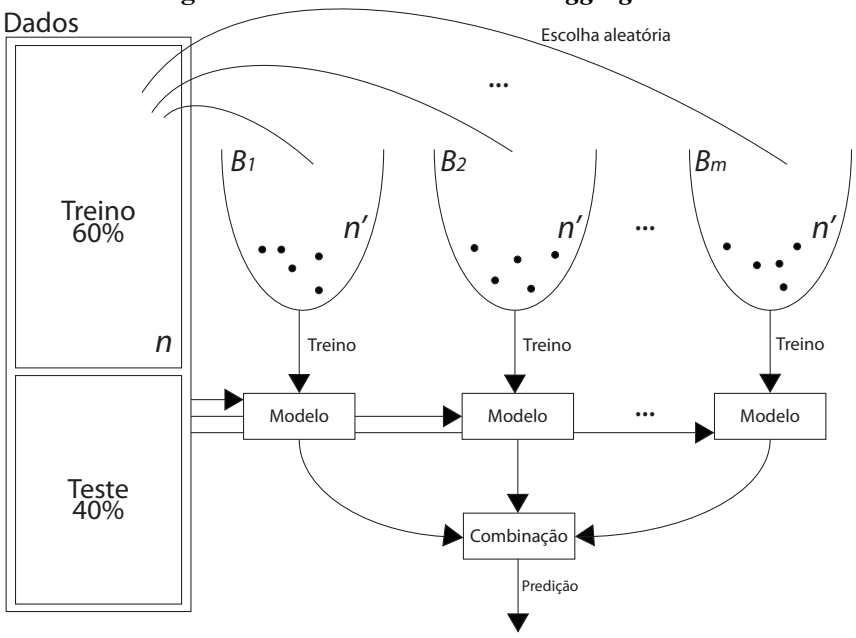
O *Bagging* (acrônimo de ***B**ootstrap **Agg**regat**ing***) foi proposto por Breiman (1996). Como o próprio sugere, esse *ensemble* é aplicado em duas etapas, inicialização (*Bootstrap*) e agregação (*Aggregating*). A inicialização consiste em criar m subdivisões aleatórias da base de dados para treinar um número m de classificadores. Dessa maneira, o foco desse *ensemble* está nos dados que são utilizados no aprendizado e não nos modelos que são gerados. Já a agregação consiste em combinar as predições dos modelos treinados a fim de realizar a predição final da amostra de teste.

O funcionamento do *Bagging* é ilustrado na Figura 15. As amostras da base são separadas em treino e teste. Com as n amostras de treino são criadas m subdivisões com n' amostras cada. Cada amostra da base é escolhida aleatoriamente e, portanto, a mesma amostra pode aparecer em mais de uma subdivisão. Ressalta-se que $n' \leq n$ em todas as subdivisões. Após a criação aleatória de cada subdivisão, os m modelos são treinados. Por fim, as amostras de teste são classificadas por todos os modelos e a predição de cada amostra pode ser obtida por votação majoritária ou média dos votos dos modelos.

O *Decision Stump*, proposto por Iba e Langley (1992), é uma árvore de decisão composta por apenas um nível. Dessa forma, esse modelo é capaz de analisar apenas uma característica do problema abordado. Por esse motivo, ele pode ser chamado de um classificador fraco. Embora não seja capaz de atingir resultados individuais satisfatórios, seu uso em *ensemble* é comum devido a sua pouca complexidade e consequente classificação rápida.

Considere uma base de dados genérica com três amostras divididas em duas classes,

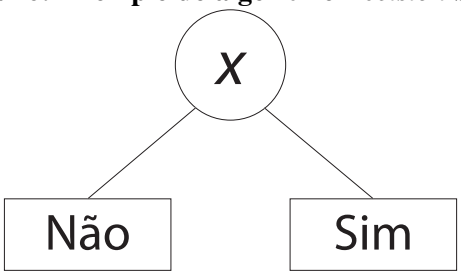
Figura 15: Funcionamento do *Bagging*.



Fonte: *https://www.youtube.com/watch?v=2Mg8QD0F1dQ* (2016)

"Sim" e "Não". Cada amostra é representada pelas características *x*, *y* e *z*. No caso do *Decision Stump*, somente uma dessas características é utilizada para classificar as amostras. Supondo ser *x* a característica que melhor classifica as amostras da base, na Figura 16 é ilustrada a estrutura da árvore correspondente.

Figura 16: Exemplo do algoritmo *Decision Stump*.



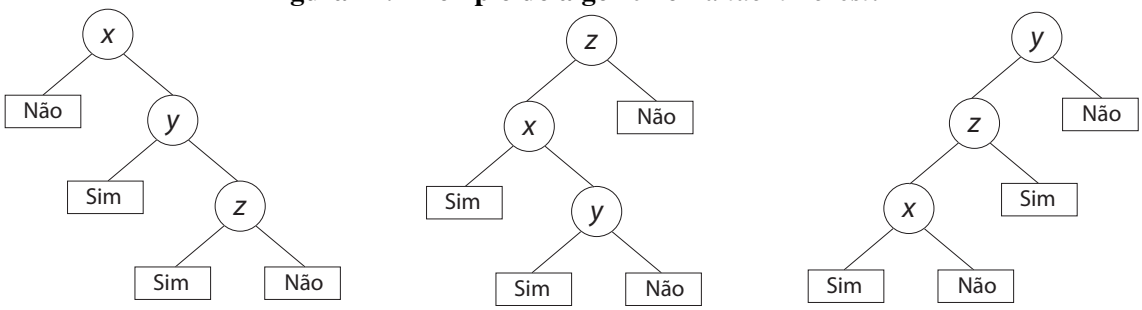
Fonte: Autoria Própria (2020)

O *Random Forest* proposto por Ho (1995), ao contrário do *Decision Stump*, é composto por diversas árvores de decisão sem restrições de tamanho e é composto por duas etapas. Na primeira etapa, as árvores de decisão são geradas levando em consideração as amostras da base bem como suas características. Para cada nó da árvore, uma característica de um determinado subconjunto aleatório de características é escolhida. É válido ressaltar que esse subconjunto é criado a partir de características que ainda não foram escolhidas na ramificação em questão. Consequentemente, é possível identificar no final do modelo uma grande variedade de árvores

criadas. Na segunda etapa, as amostras a serem classificadas são analisadas em todas as árvores e, então, os votos das árvores geradas são somados. Por conter essas duas etapas, o *Random Forest* é considerado um *Bagging*.

Considerando o mesmo exemplo enunciado na explicação do *Decision Stump*, algumas das possíveis árvores geradas no *Random Forest* são apresentadas na Figura 17.

Figura 17: Exemplo do algoritmo *Random Forest*.



Fonte: Autoria Própria (2020)

Assim como o *Random Forest*, o *Random Tree* usa da escolha aleatória de características para cada nó da árvore. Diferentemente *Random Forest*, o *Random Tree* gera apenas uma árvore de decisão e não realiza poda.

O *REPTree* (QUINLAN, 1987), acrônimo de *Reduced Error Pruning Tree*, é um modelo que cria várias árvores de decisão. Porém, diferentemente do *Random Forest*, apenas uma árvore é escolhida como a melhor e usada para a classificação (KALMEGH, 2015). A árvore é construída a partir do ganho de informação observado no cálculo da entropia, assim como as demais árvores de decisão. Após a construção, a árvore passa por um processo de poda que visa diminuir o erro resultante. Essa poda é realizada, porque segundo Witten et al. (2005), as árvores completas contém estruturas desnecessárias e é recomendado reduzi-las antes de serem implantadas.

3 TRABALHOS RELACIONADOS

É possível encontrar na literatura diversos trabalhos que abordam a tarefa de reconhecimento de emoções em músicas Music Emotion Recognition (MER). Tais trabalhos utilizam, principalmente, letras e áudios das músicas como fontes de informação para extração de características. Dessa forma, a busca por trabalhos relacionados foi guiada e organizada em termos do uso dessas duas fontes de informação.

Quanto às letras, este trabalho apresenta uma revisão sistemática na qual foram levantadas quais características, classificadores e modelos de emoção são empregados na tarefa de MER. O protocolo dessa revisão pode ser encontrado no Apêndice A e os resultados obtidos são apresentados na Subseção 3.1. Outros dois trabalhos, que não foram recuperados como parte da revisão sistemática, mas que foram importantes para o desenvolvimento deste trabalho, estão descritos na Subseção 3.2.

Em relação aos áudios, foram aproveitados alguns trabalhos da revisão sistemática, já que esses adotaram uma abordagem multimodal. Também foram recuperados trabalhos da *International Society for Music Information Retrieval* (ISMIR) de 2018, já que era a edição mais atual da conferência quando a busca foi realizada. Também fizeram parte dos estudos trabalhos que abordaram a tarefa de classificação de gêneros a fim de investigar as características derivadas dos sinais de áudio que pudessem ser utilizadas neste trabalho. Todos esses trabalhos estão descritos na Subseção 3.3.

3.1 REVISÃO SISTEMÁTICA

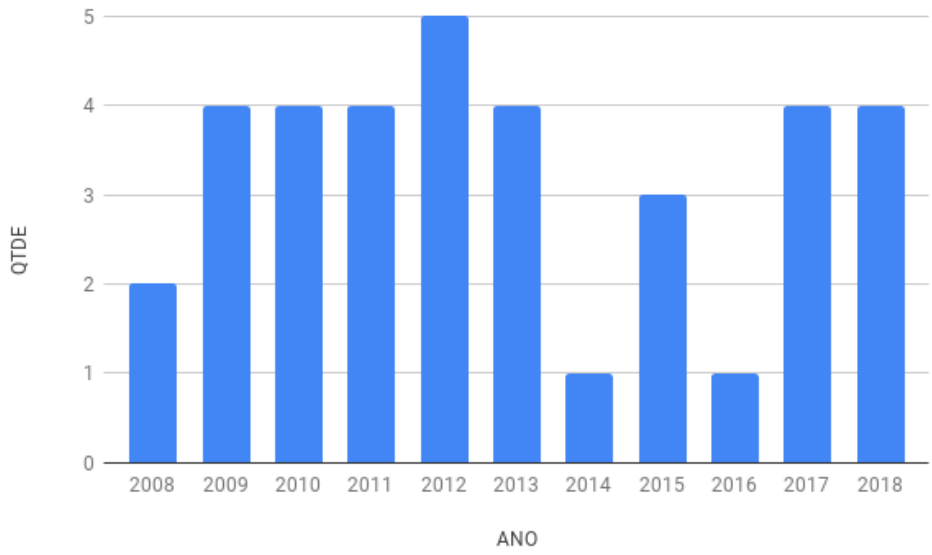
Esta subseção apresenta os resultados da análise dos 35 artigos que foram aprovados na revisão sistemática. Para melhor organização, os resultados foram separados em tópicos de acordo com a ordem definida no protocolo apresentado no Apêndice A.

3.1.1 DADOS ESTATÍSTICOS

O gráfico da Figura 18 apresenta a distribuição dos trabalhos aprovados por ano. Todos os 35 artigos foram publicados nos últimos dez anos. É possível notar também uma recuperação

na quantidade de referências no ano de 2017 em relação ao ano de 2013 após uma queda nos anos de 2014 a 2016. Esses dois fatores indicam que área tem recebido atenção da comunidade científica.

Figura 18: Trabalhos científicos por ano



Fonte: Autoria Própria (2019)

3.1.2 BASE DE DADOS

Para construir as bases de dados os autores recuperam as letras das músicas de diversas fontes encontradas na *web*. Essa recuperação aconteceu por meio de *crawlers* ou de maneira manual. As fontes utilizadas pelos trabalhos aprovados para a recuperação de letras foram: *last.fm*, *lyrics.com*, *chartLyrics.com*, *maxiLyrics.com Million Song dataset*, *azlyrics.com*, *LyricsDB*, *LyricWiki*, *lyricwiki.org*, *www.lyricsmode.com*, *xikao.com* e *www.crayonroom.com*.

Quanto a quantidade de músicas empregadas, verificou-se uma ampla variação. Dessa forma, para uma melhor visualização, os trabalhos foram agrupados em faixas de acordo com a quantidade de músicas, conforme mostra o Quadro 3.

Quadro 3: Referências divididas por faixa de quantidade de músicas

Faixa	Referências
0 a 100	Furuya et al. (2015), Chen e Tang (2018), Mihalcea e Strapparava (2012), Strapparava et al. (2012)
101 a 500	Malheiro et al. (2018), Hu e Ogihara (2012), Foucard et al. (2013), Jareanpon et al. (2018), Xiong et al. (2017), Patra et al. (2015), Lu et al. (2010), Wang et al. (2011)
501 a 1000	Patra et al. (2018), Wang et al. (2013), Kim e Kwon (2011), Laurier et al. (2008), Hu et al. (2009b), Watson e Mandryk (2012)
1001 a 5000	Ujlambkar e Attar (2012), Wu et al. (2014), Chauhan e Chauhan (2016), An et al. (2017), Su e Fung (2013), Yang e Lee (2009), Su et al. (2013), Li et al. (2015), Xia et al. (2008), Schuller et al. (2011), Zhang et al. (2017)
5001 a 10000	Hu e Downie (2010a), Dang e Shirai (2009), Hu et al. (2009a), Zaanen e Kanters (2010), Hu e Downie (2010b)
acima de 10000	McVicar et al. (2011)

Fonte: Autoria Própria (2019)

É possível destacar na faixa de 0 a 100 músicas os dois trabalhos com menor quantidade de músicas na base, Furuya et al. (2015) e Chen e Tang (2018) que utilizaram 60 músicas. Como pode ser visto na Quadro 3, a faixa de 1001 a 5000 músicas contém o maior número de trabalhos. Cinco trabalhos apresentaram uma base de dados com uma quantidade de músicas entre 5000 a 10000 e apenas o trabalho de McVicar et al. (2011) apresentou uma base de dados com uma quantidade de músicas acima de 10.000, especificamente 119.664 músicas. No entanto, cabe destacar que as músicas da base de McVicar et al. (2011) não estão rotuladas com suas respectivas emoções e foram usadas em uma tarefa de agrupamento. O tamanho relativamente pequeno das bases de dados apresentadas pode ser justificado pela dificuldade da rotulação de emoções, uma vez que essa tarefa é realizada de maneira manual, demandando tempo e recursos financeiros.

Foram coletados também dados relacionados às línguas em que as letras estão escritas. 20 trabalhos analisaram letras unicamente na língua inglesa; oito trabalhos analisaram letras unicamente na língua chinesa; dois trabalhos analisaram letras unicamente na língua indiana; um trabalho analisou letras coreanas; e um trabalho analisou letras tailandesas. Além desses,

dois trabalhos analisaram letras em mais de uma língua (por exemplo, inglês e chinês na mesma base de dados). Não foram encontrados trabalhos que analisassem letras de músicas latinas, como é o caso deste trabalho. Por fim, um trabalho não identificou a língua das letras utilizadas.

Em relação aos gêneros musicais, não foram encontrados padrões significativos que pudessem contribuir com a tarefa de classificação. Dos 35 trabalhos aprovados, apenas 10 trabalhos citaram gêneros específicos, como: *western*, hindi, rock, rap, pop, opera e músicas eletrônicas. O restante dos trabalhos indicaram apenas que as músicas eram de diversos gêneros, porém sem especifica-los.

3.1.3 MODELOS DE EMOÇÃO

Geralmente os fornecedores das letras das músicas não possuem anotações relacionadas às emoções associadas a essas músicas. Por isso, os trabalhos que optaram por uma abordagem de aprendizagem supervisionada contaram com a ajuda de *experts* para realizar a tarefa de rotulação de forma manual.

Uma outra forma de fazer a rotulação é utilizar *tags* fornecidas por usuários, como em Hu e Ogihara (2012) e Watson e Mandryk (2012). Ainda, cinco trabalhos utilizaram uma abordagem não supervisionada baseada em agrupamento.

No que diz respeito às categorias utilizadas para classificar as músicas, foram encontradas diversas abordagens, que podem ser agrupadas como segue:

- Classe A (categorias binárias): trabalhos dessa classe utilizaram uma rotulação binária. Assim, as músicas podem ser rotuladas em positiva ou negativa (XIA et al., 2008), ou ainda em alegria ou tristeza (CHAUHAN; CHAUHAN, 2016).

- Classe B (Categorias básicas de emoções): os trabalhos dessa classe utilizaram um conjunto de emoções pré-definidas, como amor, paixão, tristeza, etc.

- Classe C (Categorias genéricas de emoções): os trabalhos dessa classe utilizaram algum agrupamento de emoções, abstraindo as emoções em classes mais genéricas. Por exemplo, em Patra et al. (2018), as emoções *happy*, *delighted* e *pleased* foram agrupadas em um grupo denominado Class_Ha.

- Classe D (Categorias por quadrante): os trabalhos dessa classe utilizam quadrantes de modelos psicológicos como classes para as músicas. Os modelos mais comuns encontrados na literatura são os de Russell (1980) e o de Thayer (1990).

- Classe E (Valores de valência e excitação): os trabalhos dessa classe apresentam abordagens que se baseiam em valores de valência e excitação, e portanto, não definem classes para o aprendizado.

No Quadro 4 é apresentada uma divisão das referências analisadas nas classes descritas.

Quadro 4: Referências divididas por classes

Classe	Referências
Classe A	Chauhan e Chauhan (2016), Xia et al. (2008), Li et al. (2015)
Classe B	Ujlambkar e Attar (2012), Furuya et al. (2015), Kim e Kwon (2011), Su e Fung (2013), Yang e Lee (2009), Su et al. (2013), Foucard et al. (2013), Jareanpon et al. (2018), Xiong et al. (2017), Laurier et al. (2008), Mihalcea e Strapparava (2012), Strapparava et al. (2012), Hu et al. (2009a), Lu et al. (2010), Zaanen e Kanters (2010), Hu e Downie (2010b), Wang et al. (2011), Wu et al. (2014), Zhang et al. (2017)
Classe C	Patra et al. (2018), Chen e Tang (2018), An et al. (2017), Dang e Shirai (2009), Patra et al. (2015)
Classe D	Furuya et al. (2015), Malheiro et al. (2018), Hu et al. (2009b)
Classe E	Hu e Ogihara (2012), Wang et al. (2013), Watson e Mandryk (2012), Schuller et al. (2011), McVicar et al. (2011)

Fonte: Autoria Própria (2019)

3.1.4 CARACTERÍSTICAS E ALGORITMOS DE APRENDIZADO

Conforme especificado nos critérios de inclusão da revisão (ver protocolo disponível no Apêndice A), todos os trabalhos analisados possuem as letras das músicas como fonte de informação. Porém, 20 desses trabalhos analisaram outras fontes de informação que foram combinadas às letras das músicas. Tais trabalhos são chamados de multimodais. Destes, 17 trabalhos combinaram áudio e letras. Isso porque, como afirmam Laurier et al. (2008), essas fontes de informação podem ser complementares. Outras fontes de informação são usadas com menos frequência, como é o caso de notas musicais (dois trabalhos), meta-dados (um trabalho) e arquivos MIDI (um trabalho).

Muitas características foram exploradas pelas referências analisadas neste trabalho. É possível encontrar características convencionais, como *Bag-of-Words* (BOW) e TF-IDF, bem como características menos convencionais, como frequência de rimas e análise de semântica latente. No Quadro 5 todas as referências são sumarizadas de acordo com características exploradas.

Como pode ser observado, o uso de BOW e TF-IDF é predominante nas pesquisas, pois são relativamente simples de serem extraídos e têm boa capacidade de representação. Entretanto, é possível destacar o trabalho de Foucard et al. (2013) que fatorou a matriz de TF-IDF e assim estimulou o surgimento dos chamados tópicos. Esse processo, separa a matriz original em duas outras matrizes, sendo que uma apresenta a relevância de cada tópico para cada música e a outra descreve a contribuição de cada palavra para cada tópico. Além da ponderação com TF-IDF, em dois trabalhos, Dang e Shirai (2009) e Jareanpon et al. (2018), essa ponderação foi feita utilizando a entropia.

Quadro 5: Referências por características extraídas das letras das músicas

Característica	Referências
BOW	An et al. (2017), Wang et al. (2013), Kim e Kwon (2011), Su et al. (2013), Patra et al. (2015), Mihalcea e Strapparava (2012), Strapparava et al. (2012), Xia et al. (2008), Lu et al. (2010), Hu e Downie (2010b), Schuller et al. (2011)
BOW + TF-IDF	Patra et al. (2018), Hu e Downie (2010a), Chen e Tang (2018), Wu et al. (2014), Chauhan e Chauhan (2016), Malheiro et al. (2018), Hu e Ogihara (2012), Su e Fung (2013), Dang e Shirai (2009), Foucard et al. (2013), Jareanpon et al. (2018), Li et al. (2015), Laurier et al. (2008), Hu et al. (2009a), Zaanen e Kanters (2010), Wang et al. (2011), Zhang et al. (2017), McVicar et al. (2011),
Entropia	Dang e Shirai (2009), Jareanpon et al. (2018)
LSA	Laurier et al. (2008)
LDA	Chauhan e Chauhan (2016), Zhang et al. (2017)
Características estilísticas	Patra et al. (2018), Hu e Downie (2010a), Malheiro et al. (2018), Patra et al. (2015), Zaanen e Kanters (2010), Hu e Downie (2010b)
POS	Hu e Downie (2010a), Malheiro et al. (2018), Hu et al. (2009a), Wang et al. (2011)
Características psicolinguísticas	Patra et al. (2018), Hu e Downie (2010a), Yang e Lee (2009), Patra et al. (2015), Hu et al. (2009b), Schuller et al. (2011) Hu e Downie (2010b), Watson e Mandryk (2012), Mihalcea e Strapparava (2012), Furuya et al. (2015)
Vocabulário de palavras	Xiong et al. (2017)
Doc2Vec	Zhang et al. (2017)

Fonte: Autoria Própria (2019)

Um problema comum no uso de BOW é a alta dimensionalidade dos vetores gerados, que tendem a serem esparsos. Existem técnicas que criam espaços semânticos mais densos e que permitem diminuir essa dimensionalidade. Nesta revisão sistemática foram identificadas duas dessas técnicas baseadas em modelagem de tópicos: a LSA (Análise de Semântica Latente) e a LDA (Alocação de Dirichlet Latente). A LSA, utilizada por Laurier et al. (2008), consiste na ideia de criar um espaço semântico em que a semelhança entre os termos é dada pela ocorrência dos mesmos em contextos comuns (SOUZA, 2011). Em linhas gerais, sua aplicação parte de uma matriz composta pela frequência dos termos em cada documento. Em seguida os valores da matriz são normalizados com técnicas do tipo TF-IDF. Por fim, técnicas de álgebra linear são aplicadas para reduzir a dimensionalidade da matriz e encontrar padrões associativos. Já a LDA, utilizada por Chauhan e Chauhan (2016) e Zhang et al. (2017), é descrita por um processo generativo em que, a partir da distribuição de Dirichlet, o espaço semântico é criado com os tópicos que descrevem o documento.

As características estilísticas também são comuns, porém em todos os trabalhos encontrados elas foram utilizadas em conjunto com outras características. Destacam-se nessa categoria o uso de contagem de gírias e a frequência de rimas. A contagem de gírias, proposta por Malheiro et al. (2018), apresentou uma forte relação com músicas predominantemente negativas e com uma alta excitação. A frequência das rimas, proposta por Wang et al. (2011), apresentou bom desempenho para classificar músicas com uma baixa valência e baixa excitação.

As características extraídas de etiquetas *Part of Speech* (POS) foram utilizadas em quatro trabalhos explorados nesta revisão sistemática. Como apresentam esses trabalhos, é possível gerar *bag-of-words* baseando-se nas funções morfossintáticas das palavras em vez de usar as próprias palavras. Dessa maneira, são contados os verbos, artigos, adjetivos e assim por diante. Essa também é uma maneira de diminuir a dimensionalidade e esparsidade dos vetores gerados por meio de BOW/TF-IDF.

As características psicolinguísticas foram utilizadas em vários trabalhos obtendo resultados promissores. Léxicos de sentimentos, como *Linguistic Inquiry and Word Count* (LIWC) e *Affective Norms for English Words* (ANEW), e léxicos psicolinguísticos, como *General Inquirer* (GI), são utilizados para extrair informações de valência e excitação ou ainda categorias psicológicas ou semânticas das palavras e assim comparar as palavras desses léxicos com as palavras encontradas nas músicas.

Em vez de utilizar *bag-of-words*, Xiong et al. (2017) utilizaram as palavras das letras como rótulos que são associados aos áudios das músicas. Segundo os autores, palavras que aparecem frequentemente em músicas associadas a uma emoção específica, mas raramente em

músicas associadas a outras emoções, são geralmente mais discriminativas. Dessa maneira, foi criado um vocabulário e as palavras com maior discriminabilidade foram escolhidas.

Zhang et al. (2017) optaram por representar as músicas por meio do modelo *Paragraph Vectors*, também conhecido como Doc2Vec, que gera *embeddings* para cada documento em vez de representações convencionais como *bag-of-words*. Modelos de *embeddings* geram vetores em um espaço dimensional que é representativo de um conjunto de documentos e têm sido explorados com sucesso em diversas tarefas de processamento de linguagem natural por gerar espaços dimensionais menores e mais densos.

Por fim, foram coletados dados relacionados aos classificadores utilizados para o processo de aprendizagem. As referências foram agrupadas por classificadores e são mostrados no Quadro 6. A segunda coluna do quadro apresenta exemplos dos algoritmos pertencentes a cada grupo de classificadores.

Quadro 6: Referências por grupos de algoritmos de aprendizagem

Grupo de Classif.	**Exemplo**	**Referências**
SVM	SVM e SMO.	Patra et al. (2018), Ujlambkar e Attar (2012), Hu e Downie (2010a), Chen e Tang (2018), Malheiro et al. (2018), Kim e Kwon (2011), Su e Fung (2013), Su et al. (2013), Dang e Shirai (2009), Li et al. (2015), Xiong et al. (2017), Laurier et al. (2008), Patra et al. (2015), Strapparava et al. (2012), Xia et al. (2008), Hu et al. (2009a), Lu et al. (2010), Hu e Downie (2010b), Wang et al. (2011), Zhang et al. (2017)
Árvores de Decisão	*Random Forest*, REPTree, *decision stump*, J48 e outros.	Ujlambkar e Attar (2012), Su e Fung (2013), Yang e Lee (2009), Su et al. (2013), Jareanpon et al. (2018), Laurier et al. (2008), Wang et al. (2011), Wang et al. (2011)
Métodos Bayesianos	Naive Bayes e Bayes Net.	Ujlambkar e Attar (2012), An et al. (2017), Kim e Kwon (2011), Dang e Shirai (2009), Jareanpon et al. (2018), Wang et al. (2011), Watson e Mandryk (2012)
Métodos de Agrupamentos	Método Ward, Método Fuzzy, K-nn e K-means.	Furuya et al. (2015), Wu et al. (2014), Hu e Ogihara (2012), Laurier et al. (2008), Hu et al. (2009b), Zaanen e Kanters (2010)
Regressores	SMOReg e SVR.	Malheiro et al. (2018), Wang et al. (2013), Mihalcea e Strapparava (2012)
Outros algoritmos	MLP, HMM e outros.	Patra et al. (2018), Chauhan e Chauhan (2016), Kim e Kwon (2011), Dang e Shirai (2009), Xiong et al. (2017), Laurier et al. (2008), McVicar et al. (2011)

Fonte: Autoria Própria (2019)

Métodos de aprendizagem identificados em apenas uma referência foram agrupados na categoria "Outros Algoritmos". São eles: *Hidden Markov Model* (HMM) (KIM; KWON, 2011), *multilayer perceptron* (MLP) (PATRA et al., 2018), modelagem com grafos (DANG; SHIRAI, 2009), *topic model* (CHAUHAN; CHAUHAN, 2016), *logistic regression* (LAURIER et al., 2008), Análise de correlação canônica (CCA) (MCVICAR et al., 2011) e um método generativo multimodal elaborado por Xiong et al. (2017).

Em algumas referências foi possível notar o uso de *ensemble* de classificadores como estratégia de aprendizagem. No Quadro 7 são apresentados os *ensemble* utilizados, bem como os trabalhos em que foram aplicados.

Quadro 7: *Ensemble* encontrados

Ensemble	**Referências**
AdaBoost	Su e Fung (2013), Su et al. (2013) e Lu et al. (2010)
Bagging	Ujlambkar e Attar (2012)
DECORATE	Yang e Lee (2009)
Boosting	Foucard et al. (2013)
Generalização de Pilha	Wang et al. (2011)

Fonte: Autoria Própria (2019)

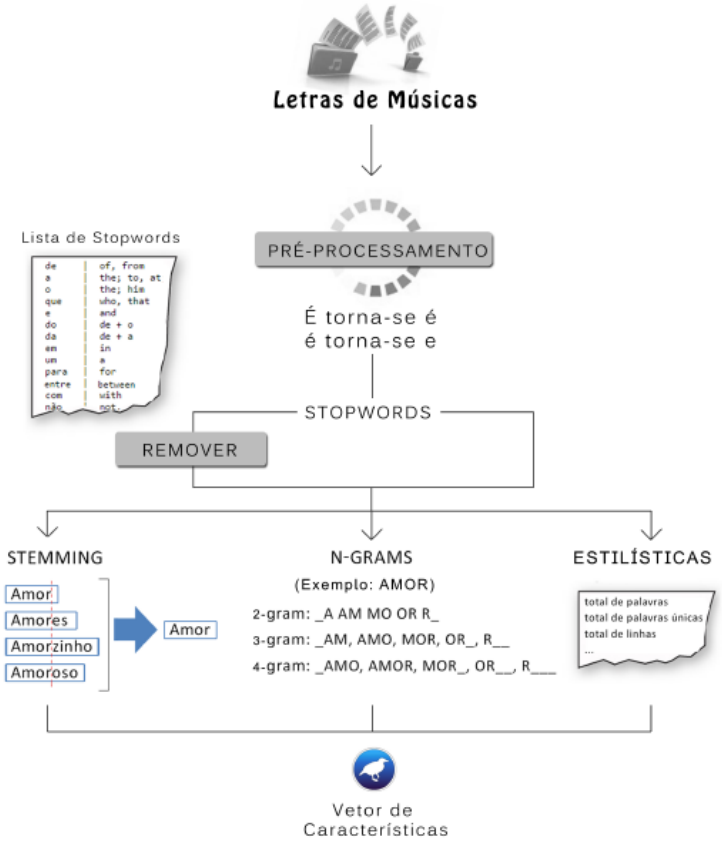
Ainda sobre os classificadores, cabe destacar que as classificações podem ser realizadas levando em consideração apenas uma emoção por música (monorrótulo) ou ainda considerando que uma música pode ser rotulada com mais de uma emoção (multirrótulo). Das 36 referências analisadas, 26 adotaram a abordagem monorrótulo, 6 referências utilizaram a abordagem multirrótulo e 4 não informaram a abordagem utilizada. É fato que uma música pode expressar mais de uma emoção, como afirma Wu et al. (2014). Porém, como constado nesta revisão sistemática, a maioria dos autores optaram por trabalhar com classificadores monorrótulo levando em consideração a emoção predominante da música ou, ainda, abordando apenas um fragmento da música.

3.2 OUTROS TRABALHOS RELACIONADOS

Dois trabalhos que dizem respeito ao uso de letras para a classificação de emoções não entraram na revisão sistemática por não estarem indexados nas bases de artigos científicos pesquisadas, porém são relevantes para o tema desta pesquisa. O primeiro, Ribeiro (2015), diz respeito ao uso de apenas letras de músicas como fonte de informação para a classificação de músicas por suas emoções e o segundo, Przybysz (2016), se utiliza de letras, áudios e acordes para realizar a mesma tarefa.

Ribeiro (2015) utilizou características estilísticas e TF-IDF a partir de n-gramas de caracteres para treinar seus modelos de classificação. Assim como em grande parte dos trabalhos relacionados à classificação de músicas por emoções baseadas em letras, o autor usou classificadores SVM. Como pré-processamentos, foram aplicados *stemming* e remoção de *stopwords* para reduzir a dimensionalidade do vocabulário das músicas e, assim, realizar experimentos mais eficientes quanto ao tempo de execução. Na Figura 19 é apresentado um esquema do extrator de características proposto pelo autor.

Figura 19: Extrator de características de Ribeiro (2015)



Fonte: Ribeiro (2015)

A partir das características extraídas, o autor utilizou duas estratégias para realizar a tarefa de classificação, uma estratégia *early fusion*, que combina as características em um único vetor, e uma estratégia *late fusion*, que combina a saída dos classificadores. Os resultados da estratégia *early fusion* foram melhores em relação à *late fusion*. Um outro ponto que o autor destaca é a remoção de *stopwords*. Pelos resultados apresentados, esse trabalho sugere que essa etapa não seja realizada. Um dos motivos que podem influenciar essa afirmação é o fato de que, como afirmam Hu e Downie (2010b), várias palavras de conteúdo são conectadas por *stopwords*, e isso faz com que a palavra tenha um significado semântico.

Przybysz (2016) classificou emoções em músicas utilizando características retiradas de três fontes de informação, a saber: letras, áudios e cifras. As características extraídas das letras são as mesmas descritas em Ribeiro (2015), exceto *TF-IDF*. Foram extraídas 161 características dos áudios por meio do descritor *Statistical Spectrum Descriptor*(SSD). Já em relação as cifras, foram utilizadas características referentes à presença de acordes, tom da música, quantidade de repetições de acordes, frequência de cada acorde e transição de acordes.

Os algoritmos de aprendizagem utilizados por Przybysz (2016) foram o SVM, K-nn e C4.5. Os resultados mostraram que o algoritmo SVM obteve o melhor desempenho. Assim como em Ribeiro (2015), foram usadas as estratégias *early fusion* e *late fusion* para combinar as diferentes fontes de informação. Nos experimentos *early fusion* as características das três fontes de informação foram combinadas em um mesmo vetor de atributos. Na estratégia *late fusion*, três classificadores foram treinados, cada um com uma fonte de informação e então regras de decisão foram aplicadas para decidir a emoção da música. No geral, os resultados dos experimentos com *early fusion* obtiveram resultados melhores que os experimentos com *late fusion*. Os resultados também mostraram um equilíbrio entre as representações sem fusão, unimodal e multimodal das características extraídas.

3.3 TRABALHOS RELACIONADOS AOS ÁUDIOS

Mesmo que a revisão sistemática tenha tido como foco trabalhos que utilizaram letras, trabalhos que também exploraram áudios e, portanto, multimodais, foram observados. Foi possível constatar que as características MFCCs (descritas na Subseção 2.2.2.2) tiveram mais atenção dos trabalhos do que as demais características. Além das características extraídas das letras, Hu e Downie (2010b) utilizaram o MARSYAS (TZANETAKIS; COOK, 2000) para extrair características acústicas. Esse *framework* extrai 63 características espectrais do áudio. Entre essas características, pode-se destacar as MFCCs, *Spectral Centroid*, *Rolloff* e *Flux*. A base utilizada pelos autores continha 5.296 músicas divididas em 18 emoções. Apesar da vasta utilização dessas características na literatura, os autores apresentaram uma análise comparativa entre as características extraídas das letras e as dos áudios e mostraram, por meio de experimentos, que características extraídas das letras superaram as extraídas dos áudios. A maior acurácia alcançada foi de 0,6172 com o uso de N-gramas, enquanto que, a acurácia de 0,5792 foi atingida com as características extraídas pelo MARSYAS.

Outro trabalho que explorou características acústicas por meio do MARSYAS foi o de Su et al. (2013). Além do MARSYAS, os *frameworks* PysSound (CABRERA et al., 1999) e openSMILE (EYBEN et al., 2010) também foram utilizados. Das letras foram extraídas apenas

N-gramas. Diferentemente do trabalho anterior, os autores combinaram as características de ambas as fontes de informação. Para essa combinação, foram utilizados os classificadores SVM e o *ensemble AdaBoost*. Para ambos os classificadores, a abordagem multimodal apresentou melhores resultados do que as abordagens unimodais. Com uma base de 3.766 músicas divididas em 14 emoções, a melhor acurácia alcançada foi de 78,19% aplicando *AdaBoost* na combinação *early fusion* de características de áudio e letras.

O trabalho de Xiong et al. (2017) também explorou as MFCC. Junto com elas, foram extraídas as características *Chromagram* que capturam aspectos relacionados a harmonia e a melodia da música, o Contraste Espectral que descreve a distribuição espectral relativa de baixo nível da música, e a Intensidade que descreve a energia total do espectro e distribuição do espectro em cada sub-banda. Das letras, foi extraído um vocabulário de *N* palavras ranqueadas de acordo com sua relevância em relação aos áudios. As características extraídas foram utilizadas para treinar um modelo *Support Vector Domain Description* (SVDD) que atingiu resultados satisfatórios (média de 62,8%). A base desse trabalho continha 4 emoções, cada uma representando um dos quadrantes do modelo de Russell. Nesse trabalho, foram utilizadas 215.000 palavras extraídas das letras das músicas e cerca de 3.300 minutos de áudio.

Outro trabalho que adotou um abordagem multimodal foi o de Patra et al. (2018). Os autores propuseram uma taxonomia de 5 grupos de emoções para rotular 319 músicas hindi e 298 músicas ocidentais e utilizaram SVM e uma Rede Neural *Feed-foward* (FFNN) como classificadores. Os principais aspectos explorados relacionados aos áudios foram intensidade, ritmo e timbre e isso foi feito por meio das ferramentas jAudio (MCKAY et al., 2005) e openSMILE (EYBEN et al., 2010). Das letras, foram extraídas N-gramas, características estilísticas e características baseadas em léxicos de sentimentos. As melhores medidas-F atingidas foram de 0,7510 para músicas *Hindi* e 0,8350 para músicas ocidentais, ambas atingidas com FFNN. Outros trabalhos que exploraram características acústicas foram os de Hu e Downie (2010a) e Ujlambkar e Attar (2012), porém com uma estrutura similar aos já descritos, tanto em relação às características, quanto a forma de classificação das músicas.

Diferentemente das abordagens comumente encontradas na literatura, nas quais se propõe alguma forma de classificação, McVicar et al. (2011) propuseram uma análise de correlação entre letras e áudios. Logo essa abordagem não foi supervisionada. Dos áudios, 65 características explorando a estrutura, a harmonia e a percussão foram extraídas com a API Echonest[1] , enquanto que, as letras foram representadas por TF-IDF. A correlação entre as características foi explorada utilizando Análise de Correlação Canônica (CCA). Como os

[1] http://developer.echonest.com/docs/v4

autores concluem, a correlação apresenta aspectos de emoções similares ao modelo de Russell, e assim descartando a necessidade de anotações manuais ou mesmo por classificadores.

Além desses, outros dois trabalhos foram recuperados a partir da base de artigos da ISMIR 2018. Delbouys et al. (2018) propuseram uma nova técnica de reconhecimento de emoções baseada em *deep learning*. A base de dados utilizada pelos autores continha 18.000 amostras de letras e áudios. Cada amostra tinha um valor contínuo de excitação e valência associado. Para predizer a valência e excitação das músicas, os autores modelaram uma Rede Neural Convolucional (CNN). Em relação aos áudios, essa rede teve como entrada os espectrogramas Mel dos áudios das músicas e deles foram extraídos vetores de tamanho 16 e 32. Para as letras, a entrada foi fornecida por uma rede *word embeddings*. Os experimentos realizados compararam o desempenho da rede com abordagens convencionais da literatura. Os resultados da CNN superaram os da literatura em relação a predição de excitação (0,2650 de R^2) e se mantiveram iguais em relação à predição de valência.

Panda et al. (2018) propuseram um conjunto de algoritmos para extrair novas características a partir da textura dos áudios. Os autores afirmam que muitas das características de áudio são de nível muito baixo, extraindo métricas abstratas do espectro ou diretamente da forma de onda de áudio e os humanos percebem conceitos musicais de nível superior, como ritmo, harmonia, melodia ou técnicas baseadas em notas, intervalos ou pontuações. Dessa forma, os autores exploraram essa hipótese e concluíram, por meio de experimentos utilizando essas características em um classificador SVM, que as novas características propostas são complementares às já existentes. A *baseline* (MARSYAS, MIR Toolbox e PsySound) dos autores atingiram 0,6750 de medida-F. Somando as novas características à *baseline* a melhor medida-F foi obtida, 0,7600.

Diferentemente dos trabalhos citados anteriormente, Tavares et al. (2017) focaram na extração de características visuais e acústicas para realizar a tarefa de classificação de músicas por emoções. Assim como em Przybysz (2016) e Ribeiro (2015), a base utilizada foi a LMMD, entretanto, com somente 1.005 amostras de áudios separadas em três emoções distintas: positiva, neutra e negativa. Foram utilizados dois tipos de descritores, a saber: os descritores acústicos e os descritores de textura. Os descritores acústicos utilizados foram o SSD (*Statistical Spectrum Descriptor*), o RP (*Rhythm Pattern*) e o RH (*Rhythm Histogram*). Com eles foi possível extrair características que computam sensações de ruídos nas 24 zonas da escala de Bark e 70 características correspondentes ao agregado de modulações de amplitude de 24 zonas críticas. Os descritores de textura utilizados foram o *Local Binary Pattern* (LBP), o *Robust Local Binary Pattern*(RLBP) e o *Local Phase Quantization*(LPQ). Com o LBP e RLBP foram

encontrados padrões locais entre os *pixels* levando em consideração a diferença de intensidade dos tons de cinza da imagem. Com o LPQ foram extraídas informações da fase local utilizando a transformada discreta de Fourier. Após a extração das características, os experimentos foram realizados com o algoritmo SVM. Inicialmente, foram modelados classificadores individuais para as características de cada descritor. Posteriormente, os classificadores foram combinados utilizando técnicas *late fusion*. Segundo os autores, as características extraídas com RLBP atingiram o melhor resultado, alcançando medida-F de 59,55%.

Os trabalhos de Tzanetakis e Cook (2002) e Bagcı e Erzin (2007), que abordaram a classificação de músicas por gênero e que utilizaram as MFCC, também serviram de base para este trabalho. Isso porque nesses artigos, as MFCC são descritas de maneira menos sucinta e, portanto, de mais fácil compreensão. Por não estarem diretamente relacionados à classificação de músicas por emoção, esses trabalhos não foram sumarizados no Quadro 2.

Para finalizar esta seção, os melhores resultados de cada trabalho recuperado na literatura foram sumarizados na Tabela 2. Nessa tabela são apresentados também, a quantidade de músicas, quantidade de emoções, características e algoritmos utilizados para atingir tais resultados. As métricas dos resultados são denotadas como: 1 - Medida-F, 2 - Precisão, 3 - Acurácia, 4 - Coeficiente de correlação, 5 - MAP, 6 - Correlação de Pearson e 7 - R^2.

Tabela 2: Sumarização dos Melhores Resultados dos Trabalhos Relacionados

Referência	Fontes de Informação	Nº de Músicas	Nº de Emoções	Características/ Frameworks	Algoritmos	Melhor Resultado
Patra et al. (2018)	Letras e Áudios	298	4	N-gramas, Estilísticas, Psicolinguísticas, jAudio, openSMILE	FFNN	0,8350[1]
Ujlambkar e Attar (2012)	Áudios	2.300	5	Diversas características acústicas incluindo MFCC	*Bagging* de *Random Forest*	0,750[2]
Hu e Downie (2010a)	Letras e Áudios	5.296	18	MARSYAS, N-gramas, FW, GI, ANEW, WordNet e Carac. Estilísticas	SVM	0,6172[3]
Furuya et al. (2015)	Letras	60	4	Tag POS	Método Ward (Clusterização)	0,3600[1]
Chen e Tang (2018)	Letras	60	4	TF-IDF	SVM	0,8890[1]
Wu et al. (2014)	Letras e Áudios	1.493	-	N-gramas, MFCC	HMER	0,1640[1]
Chauhan e Chauhan (2016)	Letras	1.900	2	N-gramas, TF-IDF	LDA	0,9000[3]
Panda et al. (2018)	Áudios	18.000	4	MARSYAS, MIR Toolbox, PsySound e Conj. de novas características	SVM	0,7600[1]
Hu e Ogihara (2012)	Letras	309	-	TF-IDF	*Framework* próprio	0,9272[1]
An et al. (2017)	Letras	3.552	3	N-gramas	*Naive Bayes*	0,6800[3]
Wang et al. (2013)	Áudios	507	Reg. Excitação	MIRtoolbox	SMOReg (Weka)	0,822[4]
Kim e Kwon (2011)	Letras	425	8	N-gramas com regras sintáticas	SVM	0,5880[3]
Su e Fung (2013)	Letras	3.766	14	N-gramas (binário)	*AdaBoost* + *Decision Stump*	0,7412[3]
Yang e Lee (2009)	Letras	1.032	23	GI	DECORATE	0,6700[3]
Su et al. (2013)	Letras e Áudios	3.766	14	N-gramas, PysSound, openSMILE, MARSYAS	*AdaBoost*	0,7819[3]

Referência	Fontes de Informação	Nº de Músicas	Nº de Emoções	Características/ Frameworks	Algoritmos	Melhor Resultado
Dang e Shirai (2009)	Letras e Metadados	6.000	5	*Artist Feature*, BOW (maior peso para títulos e refrões)	*Naive Bayes*	0,5744[3]
Foucard et al. (2013)	Letras e Áudios	500	-	MFCC, *Drum Energy*, IPP, Psysound, EchoNest, TF-IDF reduzido com Fatoração de Matriz	*Boosting* + *Decision Stump*	0,5010[5]
Jareanpon et al. (2018)	Letras	120	3	TF-IDF	*Naive Bayes* + *Decision Stump*	0,6666[3]
Li et al. (2015)	Letras	1.200	2	BAC, Bigramas, DBN	SVM	0,7380[3]
Xiong et al. (2017)	Letras e Áudios	-	4	Vocabulário por relevância, *Chromagram*, Contras. Espectral, Intensidade e MFCCs	SVDD	0,6280[3]
Laurier et al. (2008)	Letras e Áudios	1.000	4	MFCC, Centróide espectral, Ritmo, Tonal, Descritores temporais TF-IDF, LSA e termos mais discriminantes	SVM	0,9240[3]
Patra et al. (2015)	Letras	1.000	2	Estilísticas, N-gramas Léxicos de Sentimentos	SVM	0,6830[1]
Mihalcea e Strapparava (2012)	Letras e Cifras	100	6	N-gramas, LIWC, WordNet Notas e Chaves	SVM	0,5439[6]
Strapparava et al. (2012)	Letras e Cifras	100	6	N-gramas e Notas	SVM	0,7660[1]
Xia et al. (2008)	Letras	2.001	2	s-VSM	SVM	0,8849
Hu et al. (2009b)	Letras	981	4	ANCW	Método *Fuzzy* + Seleção de emoção	0,3825[1]
Hu et al. (2009a)	Áudios	5.585	18	MARSYAS	SVM	0,6151[3]

Referência	**Fontes de Informação**	**Nº de Músicas**	**Nº de Emoções**	**Características/ Frameworks**	**Algoritmos**	**Melhor Resultado**
Lu et al. (2010)	Letras, Áudios e MIDI	500	4	jSymbolic, jAudio e N-gramas	*Adaboost* + SVM	0,7240[3]
Zaanen e Kanters (2010)	Letras	5.631	2	TF, TF-IDF	TiMBL(k-NN)	0,7723[3]
Hu e Downie (2010b)	Letras	5.296	18	N-gramas	SVM	0,6172[2]
Wang et al. (2011)	Letras e Áudio	500	4	POS TF+IDF, Freq. de Ritmo	*Stacking* +SMO, Naive Bayes	0,615[1]
Schuller et al. (2011)	Letras e Áudio	2.648	-	ConceptNet, BOW, Freq. de acordes Diversas Carac. de Ritmo e *Spectral*	REPTree	0,615[4]
Watson e Mandryk (2012)	Letras e Áudio	610	2	Diversas carac. de chaves, ritmo, Articulação e Timbre, LIWC e info. contexto (ex: localização)	REPTree	0,7500[1]
Zhang et al. (2017)	Letras	1.247	-	*Paragraph Vector*	SVM	0,2580[3]
McVicar et al. (2011)	Letras e Áudios	119.664	NS	TF-IDF, Echonest API	CCA	-
Ribeiro (2015)	Letras	2.604	6	N-gramas e Estilísticas	SVM	0,5110[1]
Przybysz (2016)	Letras e Cifras	784	6	Diversas carac. de acordes, N-gramas e Estilísticas	SVM	0,3975[1]
Delbouys et al. (2018)	Letras e Áudios	18.000	2	CNN	CNN	0,2650[7]
Tavares et al. (2017)	Áudios	1.005	6	RLBP	SVM	0,5955[1]

Fonte: Autoria Própria (2020)

4 DESENVOLVIMENTO

Existem muitas bases *online* de músicas criadas por fornecedores. É possível encontrar essas bases facilmente apenas com uma rápida pesquisa na Internet ou ainda por aplicativos desenvolvidos pelos próprios fornecedores. Muitas dessas bases não possuem um sistema com uma filtragem de músicas por emoções e as que possuem, como a allmusic.com, recorrem a uma classificação manual (VALE, 2017). A partir disso, surge a necessidade de automatizar o processo de classificação de músicas. Essa foi a principal motivação deste trabalho.

São poucos os trabalhos encontrados na literatura que abordam músicas latinas. A maioria deles investigam músicas em língua inglesa, chinesa ou indiana. Portanto, uma contribuição deste trabalho foi analisar músicas escritas em espanhol e português, já que a base de dados utilizada é composta de músicas dessas duas línguas.

A *Latin Mood Music Database* (LMMD) foi utilizada para realizar os experimentos por ser, talvez, a única base com letras, áudios e músicas rotuladas com emoções disponíveis. A versão utilizada contém 2.282 músicas divididas em seis emoções, a saber: Alegria, Amor, Decepção, Excitado/Entusiasmado, Paixão e Tristeza.

A fim de realizar a tarefa de classificação, diversas características das letras e dos áudios foram extraídas. Quanto às letras, foram explorados tanto aspectos superficiais com BOW e TF-IDF, quanto aspectos semânticos com LSA e *Paragraph Vector* (D2V). A partir dos áudios, foram exploradas características extraídas diretamente do sinal de áudio (MFCC, SSD, RP, RH) e características visuais extraídas a partir do espectrograma (RLBP).

Foram utilizadas três abordagens para classificar as músicas em suas emoções, a saber: classificação simples, classificação em etapas e classificação por *ensemble*. Em todas as abordagens as características extraídas foram exploradas de forma individual e combinada.

O restante desta seção está organizado da seguinte forma: na Subseção 4.1 são apresentados os detalhes da base de dados LMMD, bem como o processo realizado para o *matching* entre as letras e os áudios da base de dados utilizada. Na Subseção 4.2 são descritas as características exploradas e como elas foram extraídas. Por fim, na Subseção 4.3, são apresentadas as três estratégias de classificação abordadas neste trabalho.

4.1 BASE LMMD E *MATCHING*

A *Latin Music Mood Database* (LMMD) é uma extensão da base *Latin Music Database* (LMD) (SILLA et al., 2008). Nela, todos os 3.139 áudios das músicas encontradas da LMD foram rotulados em seis emoções por um especialista (SANTOS; SILLA, 2015). Segundo os autores, as seis emoções foram sugeridas pelo próprio especialista que percebeu a relação das emoções escolhidas com o modelo emocional de Watson e Tellegen (1985). A partir disso, Ribeiro (2015) recuperou 2.603 letras das 3.139 músicas da LMMD utilizando a ferramenta *Ethnic Lyrics Fetcher* (RIBEIRO et al., 2014) e criou a base *Latin Music Lyrics Mood Database* (LMLMD). Assim, a LMLMD possui 2.603 letras de músicas em três línguas diferentes: inglês, espanhol e português. Essas músicas estão dividas em seis emoções, conforme é apresentado na Tabela 3.

Tabela 3: Divisão de músicas por emoções na base LMLMD.

Emoção	**N º de músicas**
Alegria	357
Amor	686
Decepção	88
Entusiasmado/Excitado	388
Paixão	708
Tristeza	376
Total	**2.603**

Fonte: Autoria Própria (2020)

Para que fosse possível explorar letras em um classificador multimodal foi necessário realizar o *matching* entre as letras da base LMLMD e os áudios da base LMMD. Nesse processo algumas músicas precisaram ser excluídas, conforme explicado a seguir.

4.1.1 *MATCHING* ENTRE LETRAS DA BASE LMLMD E AÚDIOS DA BASE LMMD

O *matching*[1] foi realizado em duas etapas. Na primeira etapa, os áudios e as letras foram relacionados de maneira automática com base no nome de cada música e seu respectivo autor. Muitas letras não continham o nome da música ou do autor, então esses dados foram completados manualmente em todas as letras das músicas. Nesse processo foram identificadas

[1] *Matching*, nesse contexto, é o processo de relacionar a letra da música com seu respectivo áudio

letras de músicas duplicadas, provavelmente devido à base LMMD conter várias versões da mesma música. A questão que se coloca é que, mesmo com versões diferentes, as letras são idênticas, o que afetaria o experimento utilizando uma mesma letra no treino e no teste. Foram identificadas 276 letras duplicadas, dividas nas seis emoções, como apresentado na Tabela 4. Todas as músicas duplicadas foram excluídas dos experimentos de modo que apenas músicas com letras únicas foram utilizadas.

Tabela 4: Quantidade de músicas duplicadas por emoção.

Emoção	**N º de músicas**
Alegria	14
Amor	74
Decepção	10
Entusiasmado/Excitado	30
Paixão	92
Tristeza	56
Total	**276**

Fonte: Autoria Própria (2020)

Também foram identificadas algumas letras em inglês. Essas músicas também foram excluídas dos experimentos, visto que a proposta foi explorar apenas músicas latinas. A quantidade de músicas identificadas com letras em inglês é apresentada na Tabela 5.

Tabela 5: Quantidade de músicas em inglês por emoção.

Emoção	**N º de músicas**
Alegria	1
Amor	3
Decepção	0
Entusiasmado/Excitado	2
Paixão	4
Tristeza	4
Total	**14**

Fonte: Autoria Própria (2020)

Alguns arquivos continham letras em português/espanhol e também a versão em inglês. Nesses arquivos, foram excluídas somente as versões em inglês.

Nessa primeira fase, várias músicas não puderam ser relacionadas aos áudios por erros de ortografia nos títulos das músicas recuperadas ou pela letra estar com o título incorreto. Assim, a segunda fase consistiu em relacionar manualmente o restante das letras das músicas com os áudios. Por fim, não foram encontrados os áudios para algumas das letras presentes na base, de modo que essas letras também foram excluídas. A quantidade dessas músicas é apresentada na Tabela 6.

Tabela 6: Quantidade de letras de músicas em que não foram encontrados os áudios.

Emoção	**N º de músicas**
Alegria	7
Amor	10
Decepção	0
Entusiasmado/Excitado	4
Paixão	8
Tristeza	2
Total	**31**

Fonte: Autoria Própria (2020)

Após o *matching* ser concluído a base ficou com 2.282 músicas. Todas as músicas possuem letra e áudio e estão distribuídas nas seis emoções, conforme apresentado na Tabela 7.

Tabela 7: Quantidade de músicas por emoções da base LMLMD utilizadas nos experimentos.

Emoção	**N º de músicas**
Alegria	335
Amor	599
Decepção	78
Entusiasmado/Excitado	350
Paixão	603
Tristeza	317
Total	**2.282**

Fonte: Autoria Própria (2020)

Como pode ser observado, a base continuou com uma distribuição similar à da base original, com a exceção das emoções Alegria e Tristeza. Assim, a base continuou desbalanceada, em especial com relação à emoção Decepção.

4.2 EXTRAÇÃO DAS CARACTERÍSTICAS

A extração de características se deu a partir das letras e dos áudios, conforme sumarizado no Quadro 8. Para os experimentos, essas características foram separadas em blocos. Os blocos 1, 2, 3 e 4 são compostos por características extraídas das letras, enquanto que os blocos 5, 6 e 7 são compostos por características extraídas dos áudios.

Quadro 8: Sumarização das Características Exploradas.

Características extraídas a partir das letras		
Bloco	**Características**	**Observação**
1	EST	Características estilísticas
2	2G (bigramas) 3G (trigramas) 4G (quadrigramas) ST (unigrama)	Características normalizadas com TF-IDF.
3	2G (bigramas) 3G (trigramas) 4G (quadrigramas) ST (unigrama)	Características normalizadas com TF-IDF e reduzidas com LSA.
4	D2V	Embeddings da rede Doc2Vec.
Características extraídas a partir dos áudios		
Bloco	**Características**	**Observação**
5	RLBP	Características texturais extraídas com RLBP.
6	TIMB	Características MFCC, Rolloff, centroide do espectro, flux, *zero crossings*.
7	SSD	Características SSD e complementares RP e RH.

Fonte: Autoria Própria (2020)

No Quadro 9 são apresentadas as 16 características estilísticas (EST) que compõem o bloco 1. Essas características foram extraídas conforme proposto nos trabalhos de Lima et al. (2014) e Przybysz (2016) e independem de língua, uma vez que estão relacionadas a contagens de palavras, linhas e caracteres especiais.

Quadro 9: Características estilísticas exploradas.

Característica	**Descrição**
Caracteres especiais de pontuação	Frequência normalizada de caracteres especiais de pontuação
Números	Frequência normalizada de números
Número de palavras	Número total de palavras
Número de palavras únicas	Número total de palavras únicas
RazaoPalavras	(No. de palavras - No. de palavras únicas) / No. de palavras
ComprimentoPalavra	Número médio de caracteres por palavras
Número de linhas	Número total de linhas
Número de linhas únicas	Número total de linhas únicas
Número de linhas em branco	Número de linhas em branco
RazaoLinhas	No. de linhas em branco / No. de linhas
ComprimentoLinha	No. de palavras / No. de linhas
DesvPadLinha	Desvio padrão do número de palavras por linhas
PalavrasUnicas	No. de palavras únicas / No. de linhas
RazaoRepetirLinha	(No. de linhas - No. de linhas únicas) / No. de linhas
PropMediaPalavrasLinha	Proporção média de palavras por linha
DesvPadPropPalavrasLinha	Desvio Padrão Prop. palavras por linha

Fonte: Lima et al. (2014)

Os n-gramas normalizados com TF-IDF também foram explorados pelos trabalhos de Przybysz (2016) e Ribeiro (2015). Assim como nos dois trabalhos citados, os n-gramas (2G, 3G e 4G) foram calculados em nível de caracteres, com a exceção do unigrama (ST), que foi calculado em nível de palavras. O pré-processamento *stemming* foi aplicado em todos os casos e não foi realizada nenhum tipo de remoção de *stopwords*, uma vez que, segundo Ribeiro (2015), esse pré-processamento piora os resultados.

A partir dos n-gramas gerados e normalizados com TF-IDF, a LSA foi aplicada para reduzir os espaços vetoriais, gerando um espaço semântico. Foram realizados experimentos com diferentes números de dimensões para encontrar um espaço que representasse bem as letras das músicas. O espaço com 100 dimensões foi o que mostrou melhores resultados.

Para calcular os vetores de características baseados em LSA foi utilizada uma técnica similar à validação cruzada. Em vez de utilizar todas as músicas para criar o espaço semântico e, desse mesmo espaço, extrair os vetores de cada música, as músicas foram divididas aleatoriamente em 50 *folds*. Assim, 49 *folds* foram utilizados para criar o espaço semântico e os vetores de características foram extraídos para o *folds* remanescente. Esse processo foi realizado para se evitar que uma mesma música tenha sido utilizada para criar o espaço semântico e dele extrair seu vetor de característica. Assim, foi possível simular um contexto mais realístico, no qual a extração de características ocorre para músicas que não se encontram na base. Tanto os vetores TF-IDF quanto os vetores LSA foram extraídas com o auxílio da biblioteca Scikit-learn (PEDREGOSA et al., 2011).

O modelo *Paragraph Vector* (D2V) também gera um espaço semântico a partir dos *embbedings* calculados em sua rede intermediária. Esses *embbedings* também foram utilizados como características no processo de classificação. Os experimentos foram realizados modificando empiricamente os parâmetros da rede e os melhores resultados foram obtidos com 250 dimensões. A métrica utilizada foi a medida-F, assim como nos demais experimentos. As configurações desses experimentos estão descritas no Apêndice B. Assim como na LSA, a técnica de *folds* foi utilizada para criação do espaço semântico e extração dos vetores de características. A ferramenta Gensim (REHUREK; SOJKA, 2011) foi utilizada para gerar os modelos D2V.

No que se diz respeito às características extraídas dos áudios, foram explorados aspectos da textura dos espectrogramas utilizando o RLBP, conforme apresentado em Tavares et al. (2017). O timbre dos áudios também foi analisado. Foram extraídas primeiramente as características MFCC e calculadas as médias de cada coeficiente para cada áudio das músicas (MFCC-med). Depois disso, além das médias, foram extraídas as variâncias das MFCC (MFCC-med+var). Por fim, foram incluídas às MFCC-var+med as médias e variâncias das características *Rolloff*, *Spectral Centroid*, *Flux*, *Zero Crossings*. Esse conjunto é denominado TIMB-var+med. Todas elas foram extraídas com a biblioteca Librosa (MCFEE et al., 2015). Por fim, foram utilizados *Rhythm Patterns* (RP), *Rhythm Histogram* (RH) e *Statistical Spectrum Descriptor* (SSD) para extrair mais características acústicas. Essas características foram extraídas com a biblioteca Rp_extract[2].

Os tamanhos dos vetores de características são apresentados no Apêndice D.

[2] github.com/tuwien-musicir/rp_extract.

4.3 ABORDAGENS DE CLASSIFICAÇÃO

Neste trabalho foram exploradas três abordagens de classificação, a saber: classificação simples, que consiste na classificação direta das seis emoções da base de dados; classificação em etapas, na qual as músicas são classificadas primeiramente em valência, excitação e quadrante para, posteriormente, serem classificadas nas emoções da base; e classificação por *ensemble*. Essas três abordagens são detalhadas a seguir.

4.3.1 CLASSIFICAÇÃO SIMPLES

Na classificação simples, as músicas foram classificadas diretamente nas seis emoções da base LMMD. Os experimentos realizados seguiram o processo apresentado na Figura 20. Os resultados de cada experimento dessa estratégia são apresentados na Subseção 5.1.

Conforme mostrado na Figura 20, as características foram avaliadas em experimentos realizados em três fases, sendo elas: fase sem combinação, fase unimodal e fase multimodal.

Na fase sem combinação as características foram avaliadas individualmente. Assim, cada característica extraída (2G, 3G, SSD, etc.) foi utilizada de forma isolada em um experimento.

Na fase unimodal as características de uma mesma fonte de informação (letra ou áudio) foram combinadas entre si. Como foram extraídas diversas características das letras, estas foram separadas em três blocos para realizar a combinação, a saber: bloco TF-IDF, bloco LSA e bloco D2V. Há também o bloco das características estilísticas, que foram combinadas com todos os outros blocos das letras.

Na fase multimodal as melhores características de cada bloco de combinações unimodal das letras foram combinadas com cada característica extraída dos áudios. Em vez de apenas combinar as melhores características de áudios com as melhores características das letras, a combinação foi realizada dessa forma para explorar melhor o espaço de combinações possíveis. A combinação foi realizada da seguinte forma: primeiramente, as características RLBP foram acrescentadas às melhores características de cada bloco. Se o resultado alcançado fosse superior ao já atingido pelas características do bloco, então as características RLBP permaneciam. Caso contrário, as características RLBP eram descartadas. O mesmo precedimento de combinação se aplicou às características TIMB e SSD, respectivamente.

Vale notar que todas as combinações, tanto unimodal quanto multimodal, foram realizadas por *early fusion*. Optou-se por não explorar combinações por *late fusion* com

base nos trabalhos de Przybysz (2016) e Ribeiro (2015). Ambos os trabalhos exploraram tais combinações com a mesma base de dados utilizada neste trabalho e obtiveram seus melhores resultados com combinações por *early fusion*. Nas três próximas subseções são apresentados os detalhes de cada estratégia de classificação implementada.

4.3.2 CLASSIFICAÇÃO EM ETAPAS

A classificação em etapas consistiu em classificar as músicas em valência, excitação, quadrante e, posteriormente, na emoção propriamente dita. Para isso, foi utilizado como base o modelo de Russell, de modo que foi necessário realizar um mapeamento das emoções da base, que são baseadas no modelo de Watson, para o modelo de Russell.

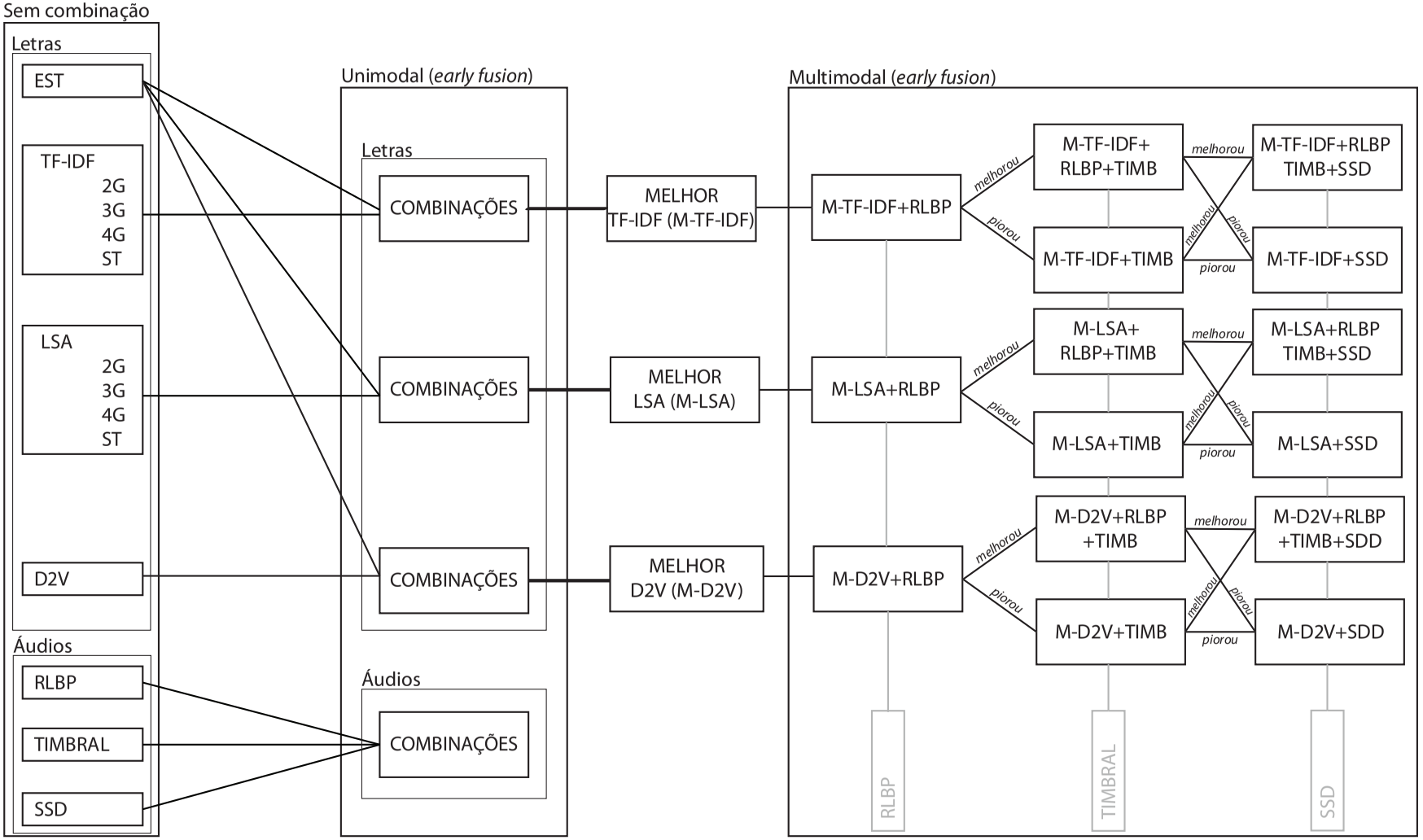
A hipótese explorada neste caso é a de que subdividir o problema da classificação de emoção em problemas menores reduziria a complexidade do problema, trazendo ganhos para o desempenho da classificação. Para encontrar o melhor classificador de valência, de excitação e de quadrante foi utilizado o mesmo processo de avaliação de características descrito para a classificação simples (Figura 20).

Tanto a classificação de valência quanto a classificação de excitação foram tratadas como problemas de classificação binária. Assim, cada música foi classificada como “positiva” ou “negativa” em relação à valência e como “alta” ou “baixa” em relação à excitação. Após o melhor classificador de valência e de excitação terem sidos encontrados, suas predições foram combinadas por meio de um conjunto de regras a fim de predizer o quadrante de cada música. Esse conjunto é formado pelas quatro regras a seguir:

1. Se a música tem valência positiva e excitação alta, então ela está no quadrante um (Q1).
2. Se a música tem valência positiva e excitação baixa, então ela está no quadrante dois (Q2).
3. Se a música tem valência negativa e excitação baixa, então ela está no quadrante três (Q3).
4. Se a música tem valência negativa e excitação alta, então ela está no quadrante quatro (Q4).

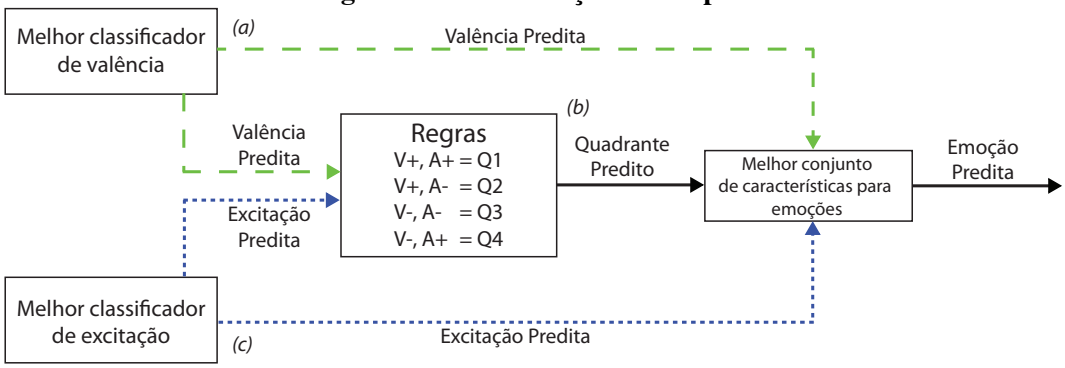
Para predizer as emoções, foram combinadas as características que obtiveram melhor resultado na primeira estratégia (classificação simples) com a predição do melhor classificador de valência, de excitação e de quadrante. Dessa forma, tais predições se tornaram características que foram concatenadas às melhores características da primeira estratégia. O processo da classificação em etapas é apresentado na Figura 21.

Figura 20: Processo de Avaliação de Características



Fonte: Autoria Própria (2020)

Figura 21: Classificação em etapas.



Fonte: Autoria Própria (2020)

4.3.2.1 MAPEAMENTO DAS EMOÇÕES

Para realizar a classificação das músicas em etapas, foi necessário classificar as músicas da base em valência, excitação e, consequentemente, em quadrante. Para isso, as emoções da base LMMD, que são baseadas no modelo de Watson e Tellegen (1985), foram mapeadas para modelo de Russell (1980). Para esse mapeamento foram levadas em consideração as descrições das emoções apresentadas em Santos e Silla (2015), bem como as explicações dos modelos de Watson e Tellegen (1985) e Russell (1980).

A emoção alegria é descrita por Santos e Silla (2015) como uma emoção primária e expressa alegrias cotidianas. Podem ser consideradas como sinônimos *acceptance* (aceitação), *satisfaction* (satisfação) e *achievement* (realização). Dessa maneira, é possível entender que essa emoção não tem um alto nível de excitação, porém é uma emoção positiva. No modelo de Watson, essa emoção representa o octante rotulado como *Low Negative Affect*. Nesse octante também estão contidas emoções como *calm* (calmo), *placid* (tranquilo) e *relaxed* (descontraído). Mapeando essa emoção para o modelo de Russell, é possível colocá-la no quadrante 2 (Q2), uma vez que, segundo o autor, esse quadrante é composto por emoções que contêm uma polaridade positiva, porém um baixo nível de excitação. Assim, a emoção alegria foi classificada como pertencendo ao quadrante 2 (Q2), com valência positiva (V+) e excitação baixa (E-).

A emoção tristeza é apresentada como *melancholy* (melancolia) e *distress* (pesar). No modelo de Watson, essa emoção representa o octante rotulado como *Unpleasantness*. No modelo de Russell, essa emoção está situada no quadrante 3 (Q3) e, portanto, pode ser descrita como tendo valência negativa (V-) e uma baixa excitação (E-).

Assim como para a emoção tristeza, a valência, excitação e o quadrante da emoção

Entusiasmado/Excitado podem ser intuitivamente deduzidos. No modelo de Watson, essa emoção representa o octante rotulado como *High Positive Affect* e é descrito por Santos e Silla (2015) como um estado de grande excitação. Logo, essa emoção foi colocada no quadrante 1 (Q1), já que esse quadrante contém emoções com valência positiva (V+) e um alto nível de excitação (E+).

A emoção decepção pode causar ambiguidade quanto ao seu nível de excitação. Da mesma forma que se pode entendê-la como uma emoção depressiva e, desse modo, classifica-la no quadrante 3 junto com a emoção tristeza, pode-se também entendê-la como frustração, o que a colocaria no quadrante 4 (Q4) de Russell. Segundo Santos e Silla (2015), o especialista que classificou as emoções das músicas da base relacionou a emoção decepção com o octante *High Negative Affect*. As emoções que compõem esse octante são emoções como *hostile* (hostil) e *nervous* (nervoso). Levando isso em consideração, a emoção decepção foi classificada como tendo valência negativa (V-) e um nível de excitação alto (E+) e, assim, foi colocada no quadrante 4 (Q4) do modelo de Russell.

A emoção amor representa as emoções que compõem o octante *Strong Engagement* do modelo de Watson (SANTOS; SILLA, 2015). Pela definição desse modelo, o fator *Disengagement-Strong Engagement* é um eixo de excitação (*arousal*). As emoções que possuem um alto engajamento, portanto, possuem uma alta excitação. Dessa maneira, a emoção amor foi colocada no quadrante 1 (Q1).

Quanto à emoção paixão, a mesma representa o octante *Pleasantness* que é composto por uma mistura de *High Postive Affect* e *Low Negative Affect*. Logo, é possível concluir que sua valência é positiva (V+). A excitação, porém, se torna difícil de classificar. Uma das emoções que está nesse octante é a emoção *happy* (alegre). Essa emoção está localizada no quadrante 1 de Russell. Porém, outra emoção que compõe esse octante é a emoção *content* (contente), localizada no quadrante 2 do modelo de Russell. Pode-se observar no modelo de Russell que a emoção paixão, propriamente dita, não é mencionada. Levando em consideração essa situação e a abordagem adotada neste trabalho, que necessita que todas as músicas tenham uma classificação binária em relação a valência e a excitação, optou-se por colocar a emoção paixão junto com as emoções amor e entusiasmado/excitado no quadrante 1 (Q1), ou seja, com alta excitação (E+).

Por fim, o mapeamento das emoções da base para o modelo de quadrantes de Russell ficou da seguinte forma:

- Alegria: valência positiva (V+), excitação negativa (E-), portanto quadrante 2 (Q2).

- Tristeza: valência negativa (V-), excitação negativa (E-), portanto quadrante 3 (Q3).
- Entusiasmado/Excitado: valência positiva (V+), excitação positiva (E+), portanto quadrante 1 (Q1).
- Decepção: valência negativa (V-), excitação positiva (E+), portanto quadrante 4 (Q4).
- Amor: valência positiva (V+), excitação positiva (E+), portanto quadrante 1 (Q1).
- Paixão: valência positiva (V+), excitação positiva (E+), portanto quadrante 1 (Q1).

4.3.3 CLASSIFICAÇÃO POR *ENSEMBLE*

Após a realização da classificação em etapas, também foram exploradas combinações por meio de dois métodos de *ensemble*, a saber: *Adaboost* e *Bagging*. Esses algoritmos foram escolhidos por terem sido utilizados por outros trabalhos da literatura com bons resultados, respectivamente, Su e Fung (2013) e Su et al. (2013), e Ujlambkar e Attar (2012). Para ambos os *ensemble*, foram utilizadas as implementações disponíveis no ambiente Weka (HALL et al., 2009).

Os algoritmos base utilizados com o *Adaboost* foram *Decision Stump*, *Random Forest*, *Random Tree* e *REPTree*. Nos trabalhos de Su e Fung (2013) e Su et al. (2013) apenas o *Decision Stump* foi explorado. Apesar dos resultados satisfatórios nesses trabalhos, quando aplicado no contexto deste trabalho, esse algoritmo base não obteve um bom desempenho. Por isso, outros algoritmos foram explorados. Já os experimentos realizados com o *Bagging* seguiram a proposta de Ujlambkar e Attar (2012) e utilizaram apenas o *Random Forest* como algoritmo base.

Foram utilizadas nos *ensemble* as melhores características de cada fase do processo de avaliação de características, sendo portanto:

- Fase sem combinação: as melhores características TF-IDF, as melhores características LSA, as melhores características D2V e as características extraídas dos áudios;
- Fase Unimodal: a melhor combinação de características TF-IDF + características estilísticas, a melhor combinação de características LSA + características estilísticas, a melhor combinação de características D2V + características estilísticas e a melhor combinação de características extraídas dos áudios;
- Fase Multimodal: a melhor combinação de características extraídas das letras e dos áudios.

Os resultados e discussões a respeito de cada uma das abordagens utilizadas neste trabalho são apresentadas na Seção 5.

5 RESULTADOS E DISCUSSÕES

Os resultados foram divididos em três subseções, sendo uma para cada estratégia de experimentos. Na Subseção 5.1 são apresentados os resultados da classificação simples, na Subseção 5.2 são apresentados os resultados da classificação em etapas e, por fim, na Subseção 5.3 são apresentados os resultados da classificação por *ensemble*.

Todos os experimentos foram realizados no ambiente do Weka com validação cruzada de 10 partições. Nas classificações simples e em etapas, o classificador *Support Vector Machine* (SVM)[1] com kernel linear e *calibrator*=SMO foi utilizado (os demais parâmetros seguiram a configuração padrão do Weka). Já na classificação por *ensemble*, o *Adaboost* e *Bagging* foram utilizados como descrito na Subseção 4.3.3. Os resultados apresentados nas tabelas a seguir correspondem às medidas-F de cada emoção, bem como a média ponderada de medida-F (coluna rotulada como Média) dos experimentos.

A medida-F, em geral obtida por uma média harmônica, é uma medida que combina valores de precisão e cobertura dos classificadores e pode ser definida como

$$Medida - F = 2 \times \frac{precisao \times cobertura}{precisao + cobertura}$$

.

5.1 CLASSIFICAÇÃO SIMPLES

Nesta subseção são apresentados os resultados dos experimentos da classificação simples. Na Subseção 5.1.1 são apresentados os resultados dos experimentos sem combinação de características, na Subseção 5.1.2 são apresentados os resultados dos experimentos unimodais e os resultados dos experimentos multimodais são apresentados na Subseção 5.1.3.

[1] o Weka implementa o *Sequential minimal optimization* (SMO), algoritmo utilizado para resolver a programação quadrática da SVM.

5.1.1 EXPERIMENTOS SEM COMBINAÇÃO

Na Tabela 8 são apresentados os resultados dos experimentos em que foram utilizadas, individualmente, as características extraídas das letras das músicas. Levando em consideração o melhor resultado dos blocos TF-IDF, LSA e D2V, pode-se observar desempenhos próximos, sendo que, o 3G (trigrama) normalizado com TF-IDF apresentou o melhor desempenho em relação as demais características. É possível notar também que as características individuais extraídas das letras apresentaram um baixo desempenho em relação às emoções com um baixo número de amostras, em especial às emoções decepção e tristeza. Esse comportamento se repetiu na maioria dos experimentos apresentados a seguir.

Tabela 8: Medida-F dos experimentos sem combinação baseados em características extraídas das letras.

Características		Alegria	Amor	Decep.	Exc./Ent.	Paixão	Tristeza	Média
EST		0,105	0,375	0,000	0,057	0,338	0,000	**0,212**
TF-IDF	**2G**	0,389	0,492	0,140	0,587	0,421	0,259	**0,428**
	3G	0,409	0,453	0,041	0,605	0,445	0,315	**0,435**
	4G	0,409	0,375	0,000	0,539	0,437	0,224	**0,387**
	ST	0,384	0,350	0,000	0,542	0,428	0,179	**0,369**
LSA	**2G**	0,263	0,469	0,000	0,520	0,350	0,127	**0,352**
	3G	0,407	0,473	0,064	0,601	0,398	0,204	**0,412**
	4G	0,314	0,477	0,000	0,621	0,408	0,225	**0,406**
	ST	0,350	0,482	0,025	0,603	0,393	0,188	**0,401**
D2V		0,427	0,489	0,074	0,590	0,402	0,303	**0,433**

Fonte: Autoria Própria (2020)

Na Tabela 9 são apresentados os resultados dos experimentos em que foram utilizadas, individualmente, as características extraídas dos áudios das músicas. Como é possível observar, os resultados das características individuais dos áudios não superaram os resultados obtidos com as características das letras. Dentre as características de áudio, destacaram-se a SSD e suas características complementares RH e RP, que obtiveram o melhor resultado.

Tabela 9: Medida-F dos experimentos sem combinação baseados em características extraídas dos áudios.

Características	Alegria	Amor	Decep.	Exc./Ent.	Paixão	Tristeza	Média
RLBP	0,133	0,443	0,025	0,461	0,247	0,036	**0,278**
MFCC-med	0,091	0,389	0,067	0,398	0,355	0,034	**0,277**
MFCC-med+var	0,144	0,438	0,126	0,465	0,284	0,073	**0,297**
TIMB-med+var	0,155	0,437	0,126	0,466	0,285	0,069	**0,298**
SSD	0,250	0,451	0,022	0,431	0,273	0,086	**0,306**
SSD+RH+RP	0,271	0,460	0,144	0,524	0,311	0,118	**0,344**

Fonte: Autoria Própria (2020)

Após esses experimentos, constatou-se que utilizar o conjunto TIMB-med+var seria melhor do que utilizar somente as MFCCs. Assim, até o final da apresentação dos resultados da classificação simples, esse conjunto de características é chamado apenas de TIMB. De maneira análoga, SSD+RH+RP obteve um melhor desempenho do que somente as características SSD. Desse modo, esse conjunto é chamado apenas de SSD deste ponto em diante.

5.1.2 EXPERIMENTOS UNIMODAIS

Na Tabela 10 são apresentados os resultados obtidos com a combinação das características estilísticas com as outras. As características TF-IDF, LSA e D2V utilizam formas diferentes de criar o espaço vetorial, e portanto, não foram combinadas entre si. Embora as características LSA e D2V criem espaços vetoriais capazes de representar informação semântica, elas não superaram a representação TF-IDF. Como se pode notar, a combinação 3G(TF-IDF)+EST obteve o melhor resultado geral. É válido ressaltar que as características estilísticas quando avaliadas sozinhas obtiveram resultados inferiores às demais características. Porém, quando combinadas com características TF-IDF e LSA, ajudaram a melhorar o resultado mesmo que de forma discreta. Analisando o contexto das características LSA e D2V, a combinação 3G+ST obteve o melhor resultado. Esse desempenho foi igualado com a combinação D2V+EST.

Tabela 10: Medida-F dos experimentos unimodais baseados em características extraídas das letras.

Características		Alegria	Amor	Decep.	Exc./Ent.	Paixão	Tristeza	Média
TF-IDF	**2G+EST**	0,411	0,489	0,119	0,607	0,413	0,259	**0,431**
	3G+EST	0,421	0,461	0,093	0,625	0,445	0,319	**0,444**
	4G+EST	0,403	0,410	0,000	0,528	0,427	0,284	**0,394**
	2G+ST	0,409	0,410	0,000	0,570	0,419	0,311	**0,409**
	3G+ST	0,371	0,419	0,000	0,604	0,443	0,293	**0,414**
	4G+ST	0,401	0,373	0,000	0,553	0,441	0,200	**0,390**
	EST+ST	0,379	0,369	0,000	0,507	0,404	0,265	**0,379**
	2G+EST+ST	0,425	0,426	0,000	0,561	0,419	0,324	**0,416**
	3G+EST+ST	0,402	0,434	0,000	0,612	0,441	0,300	**0,425**
	4G+EST+ST	0,388	0,364	0,000	0,530	0,412	0,255	**0,378**
LSA	**2G+EST**	0,324	0,457	0,130	0,534	0,347	0,131	**0,364**
	3G+EST	0,380	0,472	0,086	0,570	0,380	0,191	**0,397**
	4G+EST	0,374	0,479	0,024	0,620	0,405	0,279	**0,415**
	2G+ST	0,370	0,495	0,092	0,590	0,388	0,234	**0,413**
	3G+ST	0,384	0,510	0,063	0,611	0,403	0,253	**0,428**
	4G+ST	0,386	0,482	0,040	0,622	0,400	0,260	**0,422**
	EST+ST	0,388	0,494	0,000	0,601	0,381	0,231	**0,412**
	2G+EST+ST	0,395	0,494	0,095	0,598	0,366	0,226	**0,411**
	3G+EST+ST	0,406	0,495	0,080	0,609	0,394	0,245	**0,424**
	4G+EST+ST	0,397	0,494	0,020	0,627	0,391	0,267	**0,425**
D2V	**D2V+EST**	0,407	0,496	0,054	0,581	0,399	0,297	**0,428**

Fonte: Autoria Própria (2020)

Com relação às características extraídas dos áudios, foram realizadas todas as combinações possíveis entre características e os resultados são apresentados na Tabela 11. Como pode ser observado, a melhor combinação foi a TIMB+SSD, que superou os resultados obtidos pela SSD e suas características complementares. Os resultados das combinações das características dos áudios foram inferiores aos resultados das combinações das características das letras, mantendo o comportamento apresentado nos experimentos sem combinação. Assim, é possível concluir que as letras tiveram melhor desempenho individual em relação aos áudios na classificação de emoções da base LMMD.

Tabela 11: Medida-F dos experimentos unimodais baseados em características extraídas dos áudios.

Características	Alegria	Amor	Decep.	Exc./Ent.	Paixão	Tris.	Méd.
RLBP+TIMB	0,173	0,457	0,124	0,487	0,255	0,069	**0,301**
RLBP+SSD	0,299	0,457	0,144	0,525	0,298	0,124	**0,345**
TIMB+SSD	0,271	0,455	0,179	0,551	0,312	0,128	**0,350**
RLBP+TIMB+SSD	0,288	0,457	0,180	0,531	0,297	0,139	**0,348**

Fonte: Autoria Própria (2020)

5.1.3 EXPERIMENTOS MULTIMODAIS

Na Tabela 12 são apresentados os resultados dos experimentos multimodais, que combinaram características extraídas das letras e dos áudios. Embora as características extraídas das letras tenham apresentado melhores resultados em relação às características dos áudios, é possível observar que juntas elas atingem resultados superiores aos das letras. Isso evidencia que letras e áudios são fontes de informações complementares e que juntas auxiliam na tarefa de classificação de emoções. Devido ao grande número de combinações possíveis para as características das letras, apenas as melhores combinações das letras foram combinadas com as características dos áudios.

Tabela 12: Medida-F dos experimentos multimodais.

Características		Alegria	Amor	Decep.	Exc./Ent.	Paixão	Tristeza	Média
TF-IDF	**3G+EST+RLBP**	0,420	0,523	0,224	0,635	0,453	0,321	**0,468**
	3G+EST+RLBP +TIMB	0,418	0,516	0,211	0,643	0,457	0,339	**0,470**
	3G+EST+RLBP +TIMB+SSD	0,394	0,501	0,223	0,624	0,379	0,272	**0,431**
LSA	**3G+ST+RLBP**	0,360	0,522	0,175	0,623	0,375	0,183	**0,416**
	3G+ST+TIMB	0,399	0,519	0,235	0,629	0,388	0,237	**0,435**
	3G+ST+TIMB +SSD	0,357	0,492	0,203	0,602	0,319	0,151	**0,386**
D2V	**D2V+RLBP**	0,395	0,504	0,127	0,593	0,390	0,211	**0,418**
	D2V+TIMB	0,410	0,502	0,083	0,616	0,372	0,298	**0,429**
	D2V+SSD	0,386	0,492	0,183	0,589	0,348	0,181	**0,400**

Fonte: Autoria Própria (2020)

A combinação 3G(TF-IDF)+EST+RLBP+TIMB obteve o melhor resultado geral em todos os experimentos da classificação simples. Os resultados também mostraram que, embora as características SSD tenham um melhor desempenho em relação as outras características extraídas dos áudios, quando combinadas com as características extraídas das letras não contribuem positivamente. Esse comportamento pode ser observado nos três blocos de características das letras, TF-IDF, LSA e D2V. Vale notar também que, no geral, as características LSA tiveram desempenho inferior ao TF-IDF e ao D2V.

5.2 CLASSIFICAÇÃO EM ETAPAS

Nesta subseção são apresentados os resultados da classificação em etapas organizados da seguinte forma: na Subseção 5.2.1 são apresentados os resultados dos experimentos realizados para encontrar o melhor classificador de valência; na Subseção 5.2.2 são apresentados os resultados dos experimentos realizados para encontrar o melhor classificador de excitação; na Subseção 5.2.3 são apresentados os resultados dos experimentos realizados para encontrar o melhor classificador de quadrante; por fim, na Subseção 5.2.4, são apresentados os resultados obtidos com a combinação do melhor conjunto de características encontrado na classificação simples, 3G(TF-IDF)+EST+RLBP+TIMB, com as predições de valência, excitação e quadrante

de seus respectivos melhores classificadores.

Em relação às características MFCC e SSD extraídas dos áudios, nos experimentos de valência, excitação e quadrante notou-se o mesmo comportamento da classificação simples, em que a adição de suas respectivas características complementares melhoram os resultados. Por causa disso, nos experimentos unimodais e multimodais desta seção, TIMB refere-se as MFCCs com suas complementares e SSD refere-se à SSD+RP+RH, assim como na nos resultados da classificação simples.

5.2.1 CLASSIFICAÇÃO DE VALÊNCIA

Os experimentos realizados para avaliar as características quanto à classificação de valência seguiram o processo de combinação de características descrito na Subseção 4.3. Conforme mencionado na Subseção 4.3.2, a classificação de valência foi abordada como um problema binário. Portanto, todas as músicas foram classificadas com uma valência positiva ou negativa. É importante ressaltar que as emoções alegria, amor, excitado/entusiasmado e paixão foram classificadas com valência positiva, totalizando 1.887 músicas. Já as emoções decepção e tristeza foram rotuladas com valência negativa, totalizando 395 músicas. Desse modo, a base classificada por valência é altamente desbalanceada. A seguir são apresentados os resultados obtidos nas diferentes fases de avaliação de características.

5.2.1.1 EXPERIMENTOS SEM COMBINAÇÃO

Na Tabela 13 são apresentados os resultados das características individuais extraídas das letras. Assim como na classificação simples, a tabela é dividida em quatro blocos, cada um relacionado aos blocos das características extraídas das letras.

Tabela 13: Medida-F dos experimentos sem combinação baseados em características extraídas das letras para classificar a valência.

Características		Positiva	Negativa	Média
EST	**EST**	0,905	0,005	**0,748**
TF-IDF	**2G**	0,862	0,343	**0,772**
	3G	0,882	0,341	**0,789**
	4G	0,902	0,110	**0,765**
	ST	0,903	0,065	**0,758**
LSA	**2G**	0,870	0,190	**0,752**
	3G	0,870	0,216	**0,757**
	4G	0,877	0,209	**0,762**
	ST	0,889	0,230	**0,775**
D2V	**D2V**	0,893	0,275	**0,786**

Fonte: Autoria Própria (2020)

Na Tabela 13 é possível notar um comportamento similar ao observado na classificação simples. A classe positiva obteve resultados significativamente melhores em relação à classe negativa. Isso se deve, principalmente ao alto desbalanceamento da base, que possui um número baixo de músicas com valência negativa. Assim como na classificação por emoções, o 3G(TF-IDF) obteve o melhor resultado na classificação de valência. Destaca-se também que as características D2V apresentaram um desempenho quase tão bom quanto o observado para a 3G(TF-IDF).

A Tabela 14 apresenta os resultados obtidos com as características dos áudios. Da mesma forma que na classificação simples, as MFCC e SSD foram avaliadas separadamente e depois foram adicionadas suas características complementares. Os resultados demonstram um equilíbrio de desempenho entre as características, porém as características SSD obtiveram o melhores resultados.

Tabela 14: Medida-F dos experimentos sem fusão baseados em características extraídas dos áudios para classificar a valência.

Características	Positiva	Negativa	Média
RLBP	0,905	0,005	**0,750**
MFCC-med	0,906	0,039	**0,756**
MFCC-med+var	0,906	0,067	**0,761**
TIMB-med+var	0,906	0,072	**0,762**
SSD	0,906	0,080	**0,763**
SSD+RH+RP	0,881	0,240	**0,770**

Fonte: Autoria Própria (2020)

Comparando os resultados de 3G(TF-IDF) e SSD conclui-se que, quando avaliadas individualmente, as letras apresentam um melhor desempenho em relação aos áudios na classificação de valência. Nos dois casos (letras e áudios), os resultados obtidos para a classe positiva foram muito superiores aos da classe negativa.

5.2.1.2 EXPERIMENTOS UNIMODAIS

Na Tabela 15 são apresentados os resultados das combinações das características extraídas das letras, nas quais as características estilísticas foram combinadas com todas as outras.

Tabela 15: Medida-F dos experimentos unimodais baseados em características extraídas das letras para classificar a valência.

Características		Positiva	Negativa	Média
TF-IDF	**2G+EST**	0,858	0,338	**0,768**
	3G+EST	0,877	0,368	**0,789**
	4G+EST	0,896	0,142	**0,766**
	2G+ST	0,889	0,333	**0,793**
	3G+ST	0,892	0,242	**0,779**
	4G+ST	0,901	0,077	**0,759**
	EST+ST	0,899	0,160	**0,771**
	2G+EST+ST	0,882	0,352	**0,790**
	3G+EST+ST	0,892	0,281	**0,786**
	4G+EST+ST	0,801	0,432	**0,761**
LSA	**2G+EST**	0,869	0,182	**0,750**
	3G+EST	0,863	0,238	**0,755**
	4G+EST	0,814	0,312	**0,670**
	2G+ST	0,884	0,246	**0,774**
	3G+ST	0,895	0,239	**0,781**
	4G+ST	0,896	0,202	**0,776**
	EST+ST	0,876	0,220	**0,762**
	2G+EST+ST	0,886	0,264	**0,778**
	3G+EST+ST	0,894	0,231	**0,779**
	4G+EST+ST	0,892	0,190	**0,770**
D2V	**D2V+EST**	0,892	0,303	**0,790**

Fonte: Autoria Própria (2020)

Destacaram-se nesse conjunto de experimentos as combinações 2G(TF-IDF)+ST, 2G(TF-IDF)+ST+EST e D2V+EST, atingindo médias ponderadas de $0,793$, $0,790$ e $0,790$, respectivamente. Novamente, as características normalizadas com TF-IDF se sobressaíram em relação aos espaços semânticos criados pelas LSA e D2V.

Em relação as combinações das características extraídas dos áudios, foram realizadas todas as combinações possíveis e os resultados são apresentados na Tabela 16.

Tabela 16: Medida-F dos experimentos unimodais baseados em características extraídas dos áudios para classificar a valência.

Características	Positiva	Negativa	Média
RLBP+TIMB	0,905	0,071	**0,761**
RLBP+SSD	0,895	0,224	**0,779**
TIMB+SSD	0,887	0,224	**0,772**
RLBP+TIMB+SSD	0,894	0,209	**0,776**

Fonte: Autoria Própria (2020)

Para os áudios, a melhor combinação foi RLBP+SSD com média ponderada de 0,779, que superou o resultado de SSD alcançado nos experimentos sem combinação. Porém, esse resultado não superou o melhor resultado obtido pela combinação de características das letras, alcançado com as características 2G(TF-IDF)+ST.

5.2.1.3 EXPERIMENTOS MULTIMODAIS

Por fim, foram realizados os experimentos multimodais para a classificação de valência e seus resultados são apresentados na Tabela 17. Com relação às características TF-IDF, o melhor conjunto foi o 2G+ST. Já no contexto da LSA, o melhor conjunto foi o 3G+ST e na D2V a combinação D2V+EST foi melhor do que somente as características D2V. Desse modo, esses três conjuntos de características foram combinados com as características extraídas dos áudios.

Tabela 17: Medida-F dos experimentos multimodais para classificar a valência.

Características		Positiva	Negativa	Média
TF-IDF	**2G+ST+RLBP**	0,839	0,433	**0,769**
	2G+ST+TIMB	0,859	0,414	**0,782**
	2G+ST+SSD	0,845	0,402	**0,768**
LSA	**3G+ST+RLBP**	0,896	0,265	**0,786**
	3G+ST+RLBP+TIMB	0,896	0,268	**0,787**
	3G+ST+RLBP+TIMB+SSD	0,882	0,320	**0,785**
D2V	**D2V+EST+RLBP**	0,894	0,317	**0,794**
	D2V+EST+RLBP+TIMB	0,894	0,295	**0,790**
	D2V+EST+RLBP+SSD	0,887	0,363	**0,796**

Fonte: Autoria Própria (2020)

No contexto TF-IDF, nenhuma das combinações multimodais superou a combinação unimodal 2G+ST. No caso da LSA, todas combinações superaram o melhor resultado unimodal, obtido com 3G+ST. O melhor resultado multimodal, no entanto, foi alcançado no contexto D2V, com a combinação D2V+EST+RLBP+SSD. Essa última combinação obteve o melhor resultado geral na classificação de valência (0,796) e, portanto, foi utilizada na classificação em etapas, descrita na Subseção 5.2.4.

5.2.2 CLASSIFICAÇÃO DE EXCITAÇÃO

Os experimentos realizados para avaliar as características quanto a classificação de excitação seguiram o processo de aplicação de experimentos descrito na Subseção 4.3. A classificação de excitação também foi abordada como um problema binário. Assim, todas as músicas foram classificadas com uma excitação alta ou baixa. As emoções amor, excitado/entusiasmado, paixão e decepção foram anotadas com excitação alta totalizando 1.630 músicas. Já as emoções alegria e tristeza foram anotadas com excitação baixa totalizando 652 músicas. Desse modo, a base classificada por excitação é desbalanceada.

5.2.2.1 EXPERIMENTOS SEM COMBINAÇÃO

Os resultados das características individuais extraídas das letras são apresentados na Tabela 18. Assim como na classificação simples, a tabela é dividida em quatro blocos cada um relacionado aos blocos das características extraídas das letras.

Tabela 18: Medida-F dos experimentos sem combinação baseados em características extraídas das letras para classificar a excitação.

Características		Alta	Baixa	Média
EST	**EST**	0,883	0,000	**0,631**
TF-IDF	**2G**	0,776	0,425	**0,676**
	3G	0,759	0,479	**0,679**
	4G	0,824	0,392	**0,700**
	ST	0,828	0,202	**0,649**
LSA	**2G**	0,753	0,304	**0,624**
	3G	0,793	0,381	**0,675**
	4G	0,815	0,317	**0,673**
	ST	0,790	0,332	**0,659**
D2V	**D2V**	0,822	0,363	**0,691**

Fonte: Autoria Própria (2020)

Assim como aconteceu com a valência, para a excitação a classe alta obteve resultados significativamente melhores em relação a classe baixa devido do desbalanceamento das classes. Como pode ser notado na Tabela 18, o 4G(TF-IDF) obteve o melhor resultado na avaliação individual das características extraídas das letras para classificar a excitação. Novamente, as características extraídas pela LSA tiveram desempenho, no geral, abaixo da TF-IDF e D2V.

As características derivadas dos áudios também foram avaliadas na classificação de excitação e os resultados são apresentados na Tabela 19.

Tabela 19: Medida-F dos experimentos sem combinação baseados em características extraídas dos áudios para classificar a excitação.

Características	Alta	Baixa	Média
RLBP	0,834	0,006	**0,597**
MFCC-med	0,830	0,012	**0,596**
MFCC-med+var	0,822	0,086	**0,612**
TIMB-med+var	0,821	0,126	**0,623**
SSD	0,828	0,091	**0,617**
SSD+RH+RP	0,770	0,328	**0,664**

Fonte: Autoria Própria (2020)

Por meio da Tabela 19 é possível notar que as SSD+RH+RP alcançaram a melhor média ponderada (0,664) em relação as demais características extraídas dos áudios. Comparando os resultados de 4G e SSD+RH+RP pode-se notar que as letras se sobressaíram sobre os áudios na classificação de excitação.

5.2.2.2 EXPERIMENTOS UNIMODAIS

Na Tabela 20 são apresentados os resultados das combinações das características extraídas das letras. As características estilísticas foram combinadas com todas as outras.

Tabela 20: Medida-F dos experimentos unimodais baseados em características extraídas das letras para classificar a excitação.

Características		Alta	Baixa	Média
TF-IDF	**2G+EST**	0,781	0,414	**0,676**
	3G+EST	0,747	0,482	**0,671**
	4G+EST	0,767	0,460	**0,679**
	2G+ST	0,676	0,499	**0,625**
	3G+ST	0,767	0,490	**0,687**
	4G+ST	0,824	0,307	**0,677**
	EST+ST	0,746	0,464	**0,666**
	2G+EST+ST	0,661	0,494	**0,613**
	3G+EST+ST	0,725	0,508	**0,663**
	4G+EST+ST	0,801	0,432	**0,693**
LSA	**2G+EST**	0,764	0,350	**0,645**
	3G+EST	0,797	0,372	**0,675**
	4G+EST	0,814	0,312	**0,670**
	2G+ST	0,805	0,307	**0,663**
	3G+ST	0,811	0,293	**0,663**
	4G+ST	0,819	0,288	**0,667**
	EST+ST	0,788	0,341	**0,661**
	2G+EST+ST	0,807	0,318	**0,667**
	3G+EST+ST	0,810	0,302	**0,665**
	4G+EST+ST	0,820	0,283	**0,666**
D2V	**D2V+EST**	0,817	0,348	**0,683**

Fonte: Autoria Própria (2020)

Pode-se notar que a melhor combinação foi 4G(TF-IDF)+EST+ST(TF-IDF) atingindo $0,693$ de média ponderada. Mesmo assim, essa combinação não superou o desempenho do 4G individualmente, que foi de $0,700$. Comparando a LSA e a D2V, novamente a D2V obteve um resultado melhor.

Em relação as combinações das características extraídas dos áudios, foram realizadas todas as combinações possíveis e os resultados são apresentados na Tabela 21. O melhor resultado foi de $0,649$ alcançado com a combinação das características TIMB+SSD superando o resultado da SSD.

Tabela 21: Medida-F dos experimentos unimodais baseados em características extraídas dos áudios para classificar a Excitação.

Características	Positiva	Negativa	Média
RLBP+TIMB	0,829	0,057	**0,609**
RLBP+SSD	0,780	0,308	**0,645**
TIMB+SSD	0,770	0,347	**0,649**
RLBP+TIMB+SSD	0,779	0,308	**0,644**

Fonte: Autoria Própria (2020)

5.2.2.3 EXPERIMENTOS MULTIMODAIS

Os resultados dos experimentos multimodais são apresentados na Tabela 22. No contexto de TF-IDF, o 4G obteve o melhor resultado. Já na LSA, a melhor característica foi 3G e, no contexto do D2V, somente o *embeddings* D2V foi melhor do que a combinação D2V+EST. Assim, essas três características foram combinadas com as características extraídas dos áudios.

Tabela 22: Medida-F dos experimentos multimodais para classificar a excitação.

Características		Alta	Baixa	Média
TF-IDF	**4G+RLBP**	0,645	0,515	**0,608**
	4G+TIMB	0,666	0,507	**0,621**
	4G+SSD	0,688	0,504	**0,635**
LSA	**3G+RLBP**	0,809	0,310	**0,667**
	3G+TIMB	0,809	0,340	**0,675**
	3G+SSD	0,783	0,398	**0,673**
D2V	**D2V+RLBP**	0,817	0,364	**0,688**
	D2V+TIMB	0,820	0,366	**0,690**
	D2V+SSD	0,797	0,428	**0,692**

Fonte: Autoria Própria (2020)

No contexto TF-IDF nenhuma das combinações multimodais ou unimodais supera o resultado do 4G, que foi de 0,700. No caso da LSA, apenas a combinação 3G+TIMB chegou no mesmo resultado alcançado com o 3G. O restante das combinações multimodais foram inferiores. Para o D2V, a combinação D2V+SSD se destacou, com uma média de 0,692. Porém, esse resultado não alcançou o melhor resultado encontrado para classificação de excitação, que foi de 0,700 com o 4G. Portanto, o 4G foi utilizado na classificação em etapas descrita na Subseção 5.2.4.

5.2.3 CLASSIFICAÇÃO DE QUADRANTE

Os experimentos realizados para avaliar as características quanto a classificação de quadrante seguiram o mesmo processo de aplicação de experimentos descrito nas subseções 4.3 e 5.2.2. A classificação de quadrante foi modelada como um problema multiclasse, com uma classe representando cada um dos quatro quadrantes (Q1, Q2, Q3 e Q4). Considerando as classificações da base por quadrante, as emoções amor, Excitado/Entusiasmado e paixão foram anotadas com Q1, totalizando 1.552 músicas; a emoção alegria foi anotada com Q2, totalizando 335 músicas; a emoção tristeza foi anotada com Q3, totalizando 317 músicas; e a emoção decepção foi anotada com Q3, totalizando 78 músicas. As tabelas de resultados foram separadas como nos resultados da estratégia simples.

5.2.3.1 EXPERIMENTOS SEM COMBINAÇÃO

Na Tabela 23 são apresentados os resultados das características individuais extraídas das letras. A tabela é dividida em quatro blocos cada um relacionado aos blocos das características extraídas das letras.

Tabela 23: Medida-F dos experimentos sem combinação baseados em características extraídas das letras para classificar o quadrante.

Características		Um	Dois	Três	Quatro	Média
EST	**EST**	0,810	0,000	0,000	0,000	**0,551**
TF-IDF	**2G**	0,754	0,391	0,275	0,103	**0,612**
	3G	0,794	0,403	0,294	0,000	**0,640**
	4G	0,812	0,158	0,091	0,000	**0,640**
	ST	0,810	0,057	0,042	0,000	**0,565**
LSA	**2G**	0,751	0,243	0,140	0,000	**0,581**
	3G	0,777	0,312	0,175	0,110	**0,602**
	4G	0,794	0,237	0,217	0,000	**0,604**
	ST	0,794	0,263	0,227	0,000	**0,610**
D2V	**D2V**	0,810	0,324	0,251	0,145	**0,638**

Fonte: Autoria Própria (2020)

Assim como nos demais experimentos, a EST, quando utilizadas isoladamente, não produz resultados satisfatórios. Nas caraterísticas normalizadas com TF-IDF, destaca-se o empate das características 3G e 4G. O resultado de 0,640 atingido pelas duas características foram os melhores resultados obtidos nessa etapa. Em relação a LSA e a D2V pode-se destacar o resultado de 0,638 da D2V que quase alcançou os resultados dos 3G e 4G.

Assim como na valência e na excitação, as características extraídas dos áudios também foram avaliadas na classificação de quadrantes. Os resultados individuais dessas características são apresentados na Tabela 24.

Tabela 24: Medida-F dos experimentos sem combinação baseados em características extraídas dos áudios para classificar o quadrante.

Características	Um	Dois	Três	Quatro	Média
RLBP	0,809	0,000	0,006	0,000	**0,551**
MFCC-med	0,812	0,000	0,006	0,048	**0,555**
MFCC-med+var	0,806	0,022	0,017	0,045	**0,556**
TIMB-med+var	0,807	0,047	0,033	0,068	**0,562**
SSD	0,812	0,062	0,046	0,064	**0,570**
SSD+RH+RP	0,757	0,279	0,187	0,179	**0,588**

Fonte: Autoria Própria (2020)

Comparando os resultados de letras e áudio, mais uma vez pode-se notar que, no geral, as características extraídas letras têm desempenho melhor em relação às características extraídas dos áudios. Na subseção seguinte, as combinações unimodais apresentam esse mesmo comportamento evidenciando ainda mais o melhor desempenho das letras na tarefa de classificação de emoções.

5.2.3.2 EXPERIMENTOS UNIMODAIS

Na Tabela 25 são apresentados os resultados das combinações das características extraídas das letras. As características estilísticas foram combinadas com todas as outras.

Tabela 25: Medida-F dos experimentos unimodais baseados em características extraídas das letras para classificar o quadrante.

Características		Um	Dois	Três	Quatro	Média
TF-IDF	**2G+EST**	0,759	0,397	0,254	0,097	**0,613**
	3G+EST	0,782	0,418	0,332	0,024	**0,640**
	4G+EST	0,809	0,241	0,116	0,000	**0,602**
	2G+ST	0,767	0,419	0,285	0,000	**0,623**
	3G+ST	0,807	0,322	0,199	0,000	**0,624**
	4G+ST	0,810	0,105	0,070	0,000	**0,576**
	EST+ST	0,804	0,267	0,137	0,000	**0,605**
	2G+EST+ST	0,745	0,410	0,312	0,000	**0,610**
	3G+EST+ST	0,803	0,395	0,245	0,000	**0,638**
	4G+EST+ST	0,811	0,211	0,099	0,000	**0,596**
LSA	**2G+EST**	0,753	0,274	0,148	0,022	**0,574**
	3G+EST	0,787	0,285	0,235	0,044	**0,611**
	4G+EST	0,787	0,219	0,213	0,000	**0,597**
	2G+ST	0,792	0,259	0,167	0,000	**0,600**
	3G+ST	0,804	0,255	0,213	0,097	**0,618**
	4G+ST	0,801	0,261	0,139	0,000	**0,602**
	EST+ST	0,795	0,276	0,220	0,000	**0,612**
	2G+EST+ST	0,793	0,294	0,146	0,023	**0,603**
	3G+EST+ST	0,805	0,279	0,215	0,076	**0,621**
	4G+EST+ST	0,801	0,278	0,164	0,000	**0,608**
D2V	**D2V+EST**	0,804	0,307	0,262	0,070	**0,630**

Fonte: Autoria Própria (2020)

No contexto de TF-IDF, pode-se notar que a melhor combinação foi 3G+EST atingindo $0,64$, mesmo resultado dos 3G e 4G obtido na fase sem fusão. Na LSA, a combinação 3G+EST também se destacou, mas não superou os resultados obtidos com a normalizando TF-IDF. Por fim, a combinação D2V+EST não superou o resultado de 0,638 alcançado somente com as características D2V.

Em relação as combinações das características extraídas dos áudios, foram realizadas todas as combinações possíveis e os resultados são apresentados na Tabela 26. O melhor resultado foi de $0,595$ alcançado com a combinação das características RLBP+TIMB+SSD.

Mesmo com a melhora dos resultados, as características extraídas dos áudios não ultrapassaram os resultados dos 3G, 4G e 3G+EST.

Tabela 26: Medida-F dos experimentos unimodais baseados em características extraídas dos áudios para classificar o Quadrante.

Características	Um	Dois	Três	Quatro	Média
RLBP+TIMB	0,815	0,073	0,024	0,149	**0,574**
RLBP+SSD	0,778	0,278	0,153	0,190	**0,597**
TIMB+SSD	0,754	0,292	0,159	0,174	**0,583**
RLBP+TIMB+SSD	0,771	0,302	0,145	0,182	**0,595**

Fonte: Autoria Própria (2020)

5.2.3.3 EXPERIMENTOS MULTIMODAIS

Foram realizados também os experimentos multimodais e seus resultados são apresentados na Tabela 27. No contexto de TF-IDF, houve um empate entre os 3G, 4G e a combinação 3G+EST. Por decisão de projeto, optou-se por utilizar o 3G como característica para as combinações multimodais. Já na LSA, o melhor resultado foi obtido com a combinação 3G+EST+ST e na D2V o *embeddings* D2V foi melhor do que a combinação D2V+EST. Assim, essas características foram combinadas com as características extraídas dos áudios.

Tabela 27: Medida-F dos experimentos multimodais para classificar o quadrante.

Características		Um	Dois	Três	Quatro	Média
TF-IDF	**3G+RLBP**	0,758	0,471	0,341	0,209	**0,639**
	3G+TIMB	0,756	0,447	0,355	0,171	**0,635**
	3G+SSD	0,766	0,458	0,348	0,196	**0,644**
LSA	**3G+EST+ST+RLBP**	0,798	0,314	0,272	0,132	**0,631**
	3G+EST+ST+RLBP+TIMB	0,802	0,348	0,227	0,153	**0,633**
	3G+EST+ST+RLBP+TIMB+SSD	0,776	0,392	0,300	0,208	**0,634**
D2V	**D2V+RLBP**	0,809	0,319	0,287	0,145	**0,642**
	D2V+RLBP+TIMB	0,802	0,341	0,270	0,150	**0,638**
	D2V+RLBP+SSD	0,785	0,397	0,318	0,172	**0,642**

Fonte: Autoria Própria (2020)

No contexto TF-IDF, a combinação multimodal 3G+SSD superou os resultados do

3G, 4G e 3G+EST atingindo um resultado de 0,644. No caso da LSA, todas as combinações superaram o resultado da combinação 3G+EST+ST, porém nenhuma delas igualou o resultado da combinação 3G+EST normalizado com TF-IDF. Na D2V, as combinações D2V+RLBP+SSD e D2V+RLBP se destacaram com uma média de 0,642. Portanto, a melhor combinação encontrada para a classificação de quadrante foi 3G+SSD.

5.2.3.4 APLICANDO A TÉCNICA SMOTE

Como pôde ser observado tanto nos experimentos da classificação simples, quanto nos experimentos de valência, excitação e quadrante, o desbalanceamento foi um fator que afetou diretamente os resultados dos classificadores. Nesse contexto, a técnica SMOTE foi aplicada a fim de atenuar o desbalanceamento. Na base classificada por emoções, após a aplicação do SMOTE, a classe minoritária Decepção, que continha 78 amostras, passou a ter 156 amostras. Já na base classificada por valência, que tinha a classe negativa como minoritária, passou de 395 para 790 amostras. Na base classificada por excitação, a classe minoritária Baixa, que continha 652 amostras e passou a ter 1304 amostras. Por fim, na base classificada por quadrante, a classe minoritária Quadrante 4 (Q4) passou de 78 amostras para 156 amostras.

Após esse processo, os melhores classificadores de cada uma das classificações foi aplicado novamente e os resultados são apresentados na Tabela 28. Para fins de comparação, são apresentados também nessa tabela os resultados alcançados previamente pelos classificadores sem a técnica SMOTE. Como pode ser observado, com a exceção do classificador de valência, a técnica SMOTE proporcionou uma melhora de desempenho todos os classificadores.

Tabela 28: Resultados dos melhores classificadores de quadrante, valência, excitação e emoção com e sem uso de SMOTE.

Características		**Média**
Val.	**D2V+EST+RLBP+SSD**	**0,796**
	D2V+EST+RLBP+SSD + SMOTE	**0,716**
Exc.	**4G**	**0,700**
	4G + SMOTE	**0,914**
Quad.	**3G+SSD**	**0,644**
	3G+SSD + SMOTE	**0,656**
Emoc.	**3G+EST+RLBP+TIMB**	**0,470**
	3G+EST+RLBP+TIMB + SMOTE	**0,481**

Fonte: Autoria Própria (2020)

5.2.3.5 COMBINAÇÃO DAS PREDIÇÕES DE VALÊNCIA E EXCITAÇÃO PARA CLASSIFICAR O QUADRANTE.

A fim de melhorar o resultado da classificação de quadrante, as predições dos melhores classificadores de valência (D2V+EST+RLBP sem SMOTE) e excitação (4G com SMOTE) foram combinadas por meio das regras descritas na Subseção 4.3.2. O resultado dessa combinação é apresentado na Tabela 29. Para tentar superar o problema de desbalanceamento da base, a técnica SMOTE foi utilizada. Nessa tabela também são apresentados os resultados do melhor classificador de quadrantes até então, 3G+SSD, a fim de comparar os resultados. As letras E e V representam excitação e valência, respectivamente.

Tabela 29: Medida-F da combinação dos classificadores de valência e excitação.

Características	Um	Dois	Três	Quatro	Méd.
3G+SSD	0,766	0,458	0,348	0,196	**0,644**
3G+SSD + SMOTE	0,771	0,440	0,343	0,614	**0,656**
4G(E) + D2V+EST+RLBP(V)	0,800	0,321	0,161	0,114	**0,617**
4G (E) + SMOTE + D2V+EST+RLBP (V)	0,890	0,553	0,288	0,215	**0,734**

Fonte: Autoria Própria (2020)

Pode-se observar que a combinação 3G+SSD obteve um melhor resultado que a combinação entre 4G e D2V+EST+RLBP quando não aplicado SMOTE. Porém, esse cenário muda quando a técnica é aplicada. Os resultados da combinação entre 4G e D2V+EST+RLBP melhoram consideravelmente quando o SMOTE é aplicado no 4G.

Dessa forma o melhor classificador de quadrantes encontrado neste trabalho foi a combinação do melhor classificador de excitação (4G) com SMOTE e o melhor classificador de valência (D2V+EST+RLBP) sem a SMOTE, já que a técnica não melhorou o resultado nesse caso.

5.2.4 COMBINAÇÃO DE CLASSIFICADORES QUADRANTE, VALÊNCIA, EXCITAÇÃO PARA CLASSIFICAR A EMOÇÃO

Exclusivamente nesta subseção, os conjuntos de características receberam siglas para facilitar a apresentação dos resultados. Portanto, o melhor conjunto de características de emoções encontrado na primeira estratégia (3G+EST+RLBP+TIMB) é identificado como MC-E. A combinação entre 4G com SMOTE e D2V+EST+RLBP sem SMOTE que foi o melhor classificador de quadrante é identificado como QUAD. O mesmo acontece para valência

(D2V+EST+RLBP sem SMOTE) identificado como VA e, enfim, excitação (4G com SMOTE) identificado como EX. Os resultados das combinações são apresentados na Tabela 30.

Tabela 30: Medida-F da classificação em etapas combinando quadrante, valência e excitação com o melhor classificador da classificação simples.

Características		Alegria	Amor	Decep.	Exc./Ent.	Paixão	Tristeza	Méd.
s/ Smote	**MC-E + QUAD**	0,667	0,587	0,270	0,736	0,510	0,656	**0,600**
	MC-E + VA	0,425	0,507	0,215	0,643	0,451	0,328	**0,466**
	MC-E + EX	0,664	0,588	0,269	0,738	0,508	0,651	**0,599**
c/ Smote	**MC-E + QUAD**	0,683	0,563	0,661	0,723	0,485	0,671	**0,605**
	MC-E + VA	0,440	0,486	0,622	0,645	0,430	0,344	**0,479**
	MC-E + EX	0,672	0,562	0,667	0,725	0,480	0,663	**0,601**

Fonte: Autoria Própria (2020)

Os experimentos foram realizados com e sem a aplicação da técnica SMOTE. É possível notar que, nos três casos, a técnica proporcionou um melhor desempenho do classificador, mesmo que de maneira pouco significativa. Como é apresentado na Tabela 30, a classificação em etapas superou os resultados da classificação simples. Mesmo sem a aplicação de SMOTE na classificação final, nos três casos os resultados foram superiores aos da classificação simples. Porém, é necessário observar que para a classificação em etapas atingir bons resultados, são necessários classificadores intermediários (valência, excitação e quadrante) que tenham um bom desempenho. Caso contrário, a classificação em etapas não surte efeito.

Para comprovar isso, um exemplo é apresentado. O melhor conjunto de características encontrado para a classificação de excitação foi o 4G. Se utilizarmos essas características em um classificador intermediário de excitação sem a aplicação de SMOTE, sua medida-F é de 0,700. Usando as predições desse classificador combinado com o MC-E para classificar a emoção, a medida-F da classificação de emoções seria de 0,482. Nesse caso, a melhora do resultado foi quase nula já que o resultado atingido por MC-E com aplicação de SMOTE foi de 0,481. Porém, quando aplicado SMOTE no classificador intermediário de excitação, sua medida-F foi de 0,914. Com esse classificador foi possível atingir o resultado de 0,599 apresentado na Tabela 30. Assim, pode-se concluir que o resultado final da classificação em etapas depende diretamente do desempenho dos classificadores intermediários.

5.3 CLASSIFICAÇÃO POR *ENSEMBLE*

Nesta subseção são apresentados os resultados dos experimentos da classificação por *ensemble*. Como mencionado na Subseção 4.3.3, dois *ensemble* foram explorados, *Adaboost* e *Bagging*. As características utilizadas foram as que obtiveram os melhores resultados em cada parte do processo de aplicação de experimentos da classificação simples. Para facilitar a apresentação dos resultados nas tabelas, eles foram divididos em três blocos, a saber: sem combinação, Unimodal e Multi. (Multimodal). A letra "A" representa a fonte de informação áudio e a letra "L" representa a fonte de informação letra.

Decision Stump, *Random Forest*, *Random Tree* e *REPTree* foram utilizados como algoritmos base no *Adaboost*. Na Tabela 31 são apresentados os resultados obtidos com o *ensemble Adaboost* e com o classificador base *Decision Stump*.

Tabela 31: Medida-F dos experimentos do *ensemble Adaboost* com *Decision Stump*.

Características			**Alegria**	**Amor**	**Decep.**	**Ent.**	**Paixão**	**Tristeza**	**Média**
Sem combinação	**L**	**3G**	0,000	0,462	0,000	0,410	0,100	0,000	**0,211**
		LSA(3G)	0,000	0,520	0,000	0,442	0,000	0,000	**0,204**
		D2V	0,000	0,388	0,000	0,186	0,290	0,000	**0,207**
	A	**RLBP**	0,000	0,404	0,000	0,432	0,200	0,000	**0,225**
		TIMB	0,000	0,424	0,000	0,342	0,200	0,000	**0,217**
		SSD	0,049	0,372	0,000	0,236	0,332	0,034	**0,233**
Unimodal	**L**	**3G+EST**	0,000	0,461	0,000	0,409	0,101	0,000	**0,210**
		LSA(3G+ST)	0,127	0,521	0,000	0,428	0,000	0,000	**0,221**
		D2V+EST	0,052	0,391	0,000	0,210	0,291	0,000	**0,219**
	A	**TIMB+SDD**	0,000	0,424	0,000	0,342	0,200	0,000	**0,217**
Multi.	**L+A**	**3G+EST +RLBP +TIMB**	0,000	0,461	0,000	0,409	0,101	0,000	**0,210**

Fonte: Autoria Própria (2020)

Os resultados mostram que, mesmo com uma variedade de características e suas combinações, o classificador não obteve um bom desempenho, principalmente nas emoções decepção e tristeza. Diante desses resultados, optou-se por investigar os outros algoritmos base. Na Tabela 32 são mostrados os resultados obtidos pelo *Adaboost* e com o *Random Forest* como base.

Tabela 32: Medida-F dos experimentos do *ensemble Adaboost* com *Random Forest*.

Características			Alegria	Amor	Decep.	Ent.	Paixão	Tristeza	Média
Sem combinação	L	**3G**	0,348	0,470	0,000	0,586	0,427	0,139	**0,396**
		LSA(3G)	0,358	0,434	0,000	0,574	0,387	0,080	**0,368**
		D2V	0,293	0,432	0,000	0,446	0,375	0,040	**0,329**
	A	**RLBP**	0,243	0,413	0,035	0,502	0,330	0,190	**0,336**
		TIMB	0,221	0,447	0,141	0,481	0,347	0,167	**0,343**
		SSD	0,265	0,439	0,042	0,479	0,350	0,119	**0,338**
Unimodal	L	**3G+EST**	0,371	0,495	0,000	0,585	0,425	0,110	**0,402**
		LSA(3G+ST)	0,366	0,458	0,000	0,601	0,413	0,092	**0,388**
		D2V+EST	0,235	0,418	0,000	0,466	0,390	0,088	**0,331**
	A	**TIMB+SDD**	0,241	0,443	0,096	0,499	0,358	0,126	**0,344**
Multi.	L+A	**3G+EST +RLBP +TIMB**	0,371	0,473	0,000	0,584	0,414	0,135	**0,396**

Fonte: Autoria Própria (2020)

Os resultados do *Random Forest* foram melhores do que os do *Decision Stump*. Mesmo assim, a classificação da emoção decepção não foi satisfatória. A característica TIMB obteve o melhor resultado para essa emoção com uma medida-F de apenas 0,141. Outro ponto que pode-se destacar é que, diferente da classificação simples, a combinação multimodal não se sobressaiu em relação a unimodal e sem fusão. No geral, a característica 3G normalizada com TF-IDF combinada com EST obteve melhores resultados. Na Tabela 33 são apresentados os resultados obtidos com o *ensemble Adaboost* e com o classificador base *Random Tree*.

Tabela 33: Medida-F dos experimentos do *ensemble Adaboost* com *Random Tree*.

Características			Alegria	Amor	Decep.	Ent.	Paixão	Tristeza	Média
Sem combinação	L	**3G**	0,217	0,389	0,000	0,423	0,348	0,152	**0,312**
		LSA(3G)	0,236	0,364	0,099	0,390	0,308	0,139	**0,294**
		D2V	0,182	0,327	0,023	0,251	0,278	0,132	**0,244**
	A	**RLBP**	0,232	0,386	0,149	0,417	0,297	0,157	**0,305**
		TIMB	0,206	0,401	0,084	0,390	0,332	0,164	**0,309**
		SSD	0,200	0,353	0,121	0,397	0,311	0,154	**0,291**
Unimodal	L	**3G+EST**	0,202	0,371	0,024	0,435	0,350	0,158	**0,309**
		LSA(3G+ST)	0,301	0,375	0,022	0,472	0,341	0,233	**0,338**
		D2V+EST	0,185	0,327	0,043	0,300	0,326	0,165	**0,270**
	A	**TIMB+SDD**	0,223	0,381	0,107	0,429	0,303	0,208	**0,311**
Multi.	L+A	**3G+EST +RLBP +TIMB**	0,250	0,382	0,081	0,420	0,320	0,182	**0,314**

Fonte: Autoria Própria (2020)

Assim como nos outros experimentos dessa estratégia, é possível notar um equilíbrio com, novamente, o 3G e suas combinações atingindo melhores resultados e a classificação da emoção decepção obtendo o pior desempenho em todas as características.

Por fim, o último algoritmo base utilizado com o *Adaboost* foi o *REPTree*. Os resultados são apresentados na Tabela 34.

Tabela 34: Medida-F dos experimentos do *ensemble Adaboost* com *REPTree*.

Características			Alegria	Amor	Decep.	Ent.	Paixão	Tristeza	Média
Sem combinação	**L**	**3G**	0,317	0,402	0,000	0,510	0,342	0,170	**0,344**
		LSA(3G)	0,301	0,441	0,020	0,542	0,336	0,159	**0,355**
		D2V	0,187	0,367	0,023	0,350	0,332	0,139	**0,285**
	A	**RLBP**	0,231	0,392	0,112	0,445	0,299	0,187	**0,314**
		TIMB	0,205	0,363	0,161	0,423	0,333	0,145	**0,304**
		SSD	0,182	0,391	0,070	0,422	0,305	0,147	**0,297**
Unimodal	**L**	**3G+EST**	0,331	0,416	0,082	0,529	0,357	0,218	**0,366**
		LSA(3G+ST)	0,312	0,421	0,020	0,531	0,335	0,178	**0,352**
		D2V+EST	0,233	0,358	0,024	0,365	0,294	0,119	**0,279**
	A	**TIMB+SDD**	0,228	0,402	0,110	0,460	0,329	0,176	**0,325**
Multi.	**L+A**	**3G+EST +RLBP +TIMB**	0,295	0,433	0,103	0,508	0,381	0,191	**0,365**

Fonte: Autoria Própria (2020)

Os resultados foram inferiores aos do *Random Forest* para todas as características utilizadas. Assim como nos outros experimentos do *Adaboost*, as combinações 3G+EST e 3G+EST+RLBP+TIMB alcançaram os melhores resultados, porém não ultrapassando os resultados da classificação simples.

Após o *Adaboost* ser investigado, outro *ensemble* foi aplicado a fim de comparar os resultados. O *Bagging* foi escolhido e o *Random Forest* foi escolhido como algoritmo base, seguindo Ujlambkar e Attar (2012). Os resultados desse *ensemble* são apresentados na Tabela 35.

Tabela 35: Medida-F dos experimentos do *ensemble Bagging* com *Random Forest*.

Características			Alegria	Amor	Decep.	Ent.	Paixão	Tristeza	Média
Sem combinação	L	**3G**	0,214	0,505	0,000	0,632	0,429	0,006	**0,375**
		LSA(3G)	0,289	0,473	0,000	0,586	0,386	0,006	**0,359**
		D2V	0,057	0,445	0,000	0,338	0,391	0,000	**0,280**
	A	**RLBP**	0,188	0,426	0,039	0,510	0,304	0,104	**0,314**
		TIMB	0,109	0,445	0,067	0,484	0,310	0,109	**0,306**
		SSD	0,224	0,447	0,023	0,515	0,355	0,132	**0,342**
Unimodal	L	**3G+EST**	0,197	0,479	0,000	0,619	0,387	0,012	**0,353**
		LSA(3G+ST)	0,273	0,514	0,000	0,624	0,383	0,000	**0,372**
		D2V+EST	0,099	0,455	0,000	0,354	0,366	0,000	**0,285**
	A	**TIMB+SDD**	0,222	0,464	0,022	0,531	0,351	0,121	**0,346**
Multi.	L+A	**3G+EST +RLBP +TIMB**	0,148	0,522	0,000	0,632	0,375	0,039	**0,360**

Fonte: Autoria Própria (2020)

Os resultados do *Bagging* com Random Forest não foram satisfatórios. Novamente se destaca negativamente a classificação das emoções decepção e tristeza. O melhor resultado foi obtido com a combinação de 3G+ST enquanto que a D2V individualmente obteve o pior resultado.

Comparando-se os resultados dos dois *ensemble* é possível concluir que o *Adaboost* obteve o melhor desempenho quando usado com o *Random Forest*. De forma geral, tanto os resultados da classificação simples quanto os resultados da classificação em etapas foram melhores do que os resultados da classificação por *ensemble*.

6 CONCLUSÃO

A proposta desta pesquisa foi modelar um classificador multimodal baseado em características extraídas de áudios e letras capaz de identificar a emoção de uma música. A LMMD foi utilizada como base de dados, pois contém músicas latinas que são pouco exploradas na literatura.

A partir dos resultados da revisão sistemática realizada, foram extraídas características das letras que representaram, tanto aspectos superficiais, quanto aspectos semânticos. Os aspectos superficiais foram representados por n-gramas normalizados com TF-IDF e pelas características estilísticas, enquanto que os aspectos semânticos foram representados pela LSA e D2V.

Em relação aos áudios, foram extraídas características relacionadas a textura do espectrograma com RLBP e características acústicas com MFCCs combinadas com *Rolloff*, *Spectral centroid*, *Flux* e *Zerocrossings*, SSD, RP e RH. As características das duas fontes de informação (letra e áudio) foram analisadas de forma individual e combinadas entre si.

No decorrer dos experimentos, notou-se que, individualmente, as características extraídas das letras apresentaram um melhor desempenho em relação às características extraídas dos áudios. Além disso, os experimentos mostraram que a combinação entre essas características é mais eficaz do que utilizar apenas uma fonte de informação. Isso porque, em geral, os experimentos multimodais obtiveram melhores resultados do que os experimentos sem combinação e unimodal.

De forma individual, as características TF-IDF tiveram um melhor desempenho que a LSA e D2V. Um dos motivos que pode ter levado a esse comportamento é o fato da base ter poucas amostras e os modelos LSA e D2V tendem a obter melhores resultados com um grande número de amostras. Em relação aos áudios, as características SSD+RP+RH obtiveram os melhores resultados .

Para modelar os classificadores, três abordagens foram propostas, a saber: a classificação simples, em que as músicas foram classificadas diretamente nas seis emoções da base utilizando o classificador SVM; a classificação em etapas, também utilizando o SVM, em que as músicas foram classificadas em valência, excitação e quadrante e, a partir dessas

predições, classificadas em emoções; e a classificação por *ensemble*, em que as músicas foram classificadas diretamente nas seis emoções utilizando *Adaboost* e *Bagging*.

Na classificação simples o melhor resultado foi alcançado com a combinação multimodal 3G(TF-IDF)+EST+RLBP+TIMB, atingindo uma medida-F média de $0,47$. Esse resultado foi superado na classificação em etapas. Após ser encontrado o melhor classificador de valência (medida-F média de $0,796$, atingida com a combinação D2V+EST+RLBP+SSD), e o melhor classificador de excitação (medida-F média de $0,914$, atingida com a combinação 4G(TF-IDF) + SMOTE), as predições desses classificadores foram combinadas por quatro regras a fim de se obter um melhor classificador de quadrantes. Essa combinação foi bem sucedida, pois o melhor classificador de quadrantes antes encontrado (3G(TF-IDF)+SSD + SMOTE) foi superado por essa combinação, que atingiu medida-F média de $0,734$. Por fim, as predições de quadrante foram combinadas com o melhor conjunto de características encontrado na classificação simples (3G(TF-IDF)+EST+RLBP+TIMB) para classificar as músicas por emoções. Os resultados obtidos foram de $0,600$ e $0,605$ sem e com o uso do SMOTE, respectivamente. Dessa forma a classificação em etapas superou a classificação simples evidenciando que, subdividir o problema de classificação pode ser uma estratégia melhor do que classificar diretamente as músicas por emoções.

Na classificação por *ensemble*, foram utilizadas as melhores características encontradas na classificação simples. Contudo, em geral, os resultados foram inferiores aos das classificações simples e em etapas. A melhor medida-F média alcançada foi de $0,396$, atingida pelo *Adaboost* com *Random Forest* utilizando as características 3G(TF-IDF) e 3G(TF-IDF)+EST+RLBP+TIMB.

Em todos os experimentos, notou-se que o desbalanceamento da base de dados afetou negativamente os resultados. Como pode ser observado, a aplicação da técnica SMOTE como forma de reduzir o desbalanceamento produziu melhora dos resultados. Em todos os experimentos que a técnica foi utilizada os resultados melhoraram, com exceção do melhor classificador de valência.

Como trabalhos futuros é possível explorar novas características que apareceram na revisão sistemática, mas que não foram utilizadas neste estudo, como as características psicolinguísticas. Além disso, é possível avaliar os classificadores propostos em outras bases de dados, mesmo que não de músicas latinas, uma vez que as características utilizadas podem ser ajustadas a outras línguas. Em especial, bases de dados com mais amostras podem ser exploradas para investigar se o uso dos modelos LSA e D2V mantêm resultados inferiores às características TF-IDF.

REFERÊNCIAS

ABU-MOSTAFA, Y.; LIN, H.; MAGDON-ISMAIL, M. Learning from data: a short course: Amlbook. **View Article PubMed/NCBI**, 2012.

AN, Y.; SUN, S.; WANG, S. Naive Bayes classifiers for music emotion classification based on lyrics. In: **2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)**. [S.l.: s.n.], 2017. p. 635–638.

BAGCI, U.; ERZIN, E. Automatic classification of musical genres using inter-genre similarity. **IEEE Signal Processing Letters**, v. 14, n. 8, p. 521, 2007.

BREIMAN, L. Bagging predictors. **Machine learning**, Springer, v. 24, n. 2, p. 123–140, 1996.

CABRERA, D. et al. Psysound: A computer program for psychoacoustical analysis. In: **Proceedings of the Australian Acoustical Society Conference**. [S.l.: s.n.], 1999. v. 24, p. 47–54.

CHAUHAN, S.; CHAUHAN, P. Music mood classification based on lyrical analysis of Hindi songs using Latent Dirichlet Allocation. In: **2016 International Conference on Information Technology (InCITe) - The Next Generation IT Summit on the Theme - Internet of Things: Connect your Worlds**. [S.l.: s.n.], 2016. p. 72–76.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, p. 321–357, 2002.

CHEN, J.; KELLOKUMPU, V.; ZHAO, G.; PIETIKÄINEN, M. Rlbp: Robust local binary pattern. In: **BMVC**. [S.l.: s.n.], 2013.

CHEN, X.; TANG, T. Y. Combining Content and Sentiment Analysis on Lyrics for a Lightweight Emotion-Aware Chinese Song Recommendation System. In: **Proceedings of the 2018 10th International Conference on Machine Learning and Computing**. New York, NY, USA: ACM, 2018. (ICMLC 2018), p. 85–89. ISBN 978-1-4503-6353-2. Disponível em: <http://doi.acm.org/10.1145/3195106.3195148>.

COSTA, Y. M.; OLIVEIRA, L. S.; JR, C. N. S. An evaluation of convolutional neural networks for music classification using spectrograms. **Applied soft computing**, Elsevier, v. 52, p. 28–38, 2017.

COSTA, Y. M. e Gomes da. **Reconhecimento de Gêneros Musicais Utilizando Espectrogramas com Combinação de Classificadores**. Tese (Doutorado) — Universidade Federal do Paraná, 2013.

DANG, T.; SHIRAI, K. Machine Learning Approaches for Mood Classification of Songs toward Music Search Engine. In: **2009 International Conference on Knowledge and Systems Engineering**. [S.l.: s.n.], 2009. p. 144–149.

DEERWESTER, S.; DUMAIS, S. T.; FURNAS, G. W.; LANDAUER, T. K.; HARSHMAN, R. Indexing by latent semantic analysis. **Journal of the American society for information science**, Wiley Online Library, v. 41, n. 6, p. 391–407, 1990.

DELBOUYS, R.; HENNEQUIN, R.; PICCOLI, F.; ROYO-LETELIER, J.; MOUSSALLAM, M. Music mood detection based on audio and lyrics with deep neural net. **arXiv preprint arXiv:1809.07276**, 2018.

DHANARAJ, R.; LOGAN, B. Automatic prediction of hit songs. In: **International Society for Music Information Retrieval**. [S.l.: s.n.], 2005. p. 488–491.

EYBEN, F.; WÖLLMER, M.; SCHULLER, B. Opensmile: the munich versatile and fast open-source audio feature extractor. In: ACM. **Proceedings of the 18th ACM international conference on Multimedia**. [S.l.], 2010. p. 1459–1462.

FOUCARD, R.; ESSID, S.; RICHARD, G.; LAGRANGE, M. Exploring new features for music classification. In: **2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)**. [S.l.: s.n.], 2013. p. 1–4.

FREUND, Y.; SCHAPIRE, R. E. A desicion-theoretic generalization of on-line learning and an application to boosting. In: SPRINGER. **European conference on computational learning theory**. [S.l.], 1995. p. 23–37.

FULOP, S. A. **Speech spectrum analysis**. [S.l.]: Springer Science & Business Media, 2011.

FURUYA, M.; OKU, K.; KAWAGOE, K. Music Feeling Classification Based on Lyrics Using Weighting of Non-emotional Words. In: **Proceedings of the 13th International Conference on Advances in Mobile Computing and Multimedia**. New York, NY, USA: ACM, 2015. (MoMM 2015), p. 380–383. ISBN 978-1-4503-3493-8. Disponível em: <http://doi.acm.org/10.1145/2837126.2843844>.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The weka data mining software: an update. **ACM SIGKDD explorations newsletter**, ACM, v. 11, n. 1, p. 10–18, 2009.

HO, T. K. Random decision forests. In: IEEE. **Proceedings of 3rd international conference on document analysis and recognition**. [S.l.], 1995. v. 1, p. 278–282.

HU, X.; DOWNIE, J. S. Improving Mood Classification in Music Digital Libraries by Combining Lyrics and Audio. In: **Proceedings of the 10th Annual Joint Conference on Digital Libraries**. New York, NY, USA: ACM, 2010. (JCDL '10), p. 159–168. ISBN 978-1-4503-0085-8. Disponível em: <http://doi.acm.org/10.1145/1816123.1816146>.

HU, X.; DOWNIE, J. S. When lyrics outperform audio for music mood classification: A feature analysis. In: **International Society for Music Information Retrieval**. [S.l.: s.n.], 2010. p. 619–624.

HU, X.; DOWNIE, J. S.; EHMANN, A. F. Lyric text mining in music mood classification. **American music**, v. 183, n. 5,049, p. 2–209, 2009.

HU, Y.; CHEN, X.; YANG, D. Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In: **International Society for Music Information Retrieval**. [S.l.: s.n.], 2009. p. 123–128.

HU, Y.; OGIHARA, M. Identifying Accuracy of Social Tags by Using Clustering Representations of Song Lyrics. In: **2012 11th International Conference on Machine Learning and Applications**. [S.l.: s.n.], 2012. v. 1, p. 582–585.

IBA, W.; LANGLEY, P. Induction of one-level decision trees. In: **Machine Learning Proceedings 1992**. [S.l.]: Elsevier, 1992. p. 233–240.

ITO, F. T. **REPRESENTAÇÃO MULTIMODAL PARA CLASSIFICAÇÃO DE INFORMAÇÃO**. Dissertação (Mestrado) — UNIVERSIDADE FEDERAL DE SÃO CARLOS, 2018.

JAREANPON, C.; KIATJINDARAT, W.; POLHOME, T.; KHONGKRAPHAN, K. Automatic lyrics classification system using text mining technique. In: **2018 International Workshop on Advanced Image Technology (IWAIT)**. [S.l.: s.n.], 2018. p. 1–4.

JURAFSKY, D.; MARTIN, J. **Speech and Language Processing: An Introduction to Natural Language Processing**. [S.l.]: Prentice Hall, 1999.

KALMEGH, S. Analysis of weka data mining algorithm reptree. **Simple Cart and RandomTree for Classification of Indian News**, 2015.

KIM, M.; KWON, H. Lyrics-Based Emotion Classification Using Feature Selection by Partial Syntactic Analysis. In: **2011 IEEE 23rd International Conference on Tools with Artificial Intelligence**. [S.l.: s.n.], 2011. p. 960–964.

LAURIER, C.; GRIVOLLA, J.; HERRERA, P. Multimodal Music Mood Classification Using Audio and Lyrics. In: **2008 Seventh International Conference on Machine Learning and Applications**. [S.l.: s.n.], 2008. p. 688–693.

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: **International conference on machine learning**. [S.l.: s.n.], 2014. p. 1188–1196.

LI, J.; GAO, S.; HAN, N.; FANG, Z.; LIAO, J. Music Mood Classification via Deep Belief Network. In: **2015 IEEE International Conference on Data Mining Workshop (ICDMW)**. [S.l.: s.n.], 2015. p. 1241–1245.

LIDY, T.; RAUBER, A. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In: **International Society for Music Information Retrieval**. [S.l.: s.n.], 2005. p. 34–41.

LIMA, A. A. de; NUNES, R. M.; RIBEIRO, R. P.; SILLA, C. N. Nordic music genre classification using song lyrics. In: SPRINGER. **International Conference on Applications of Natural Language to Data Bases/Information Systems**. [S.l.], 2014. p. 89–100.

LORENA, A. C.; CARVALHO, A. C. de. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43–67, 2007.

LU, Q.; CHEN, X.; YANG, D.; WANG, J. Boosting for multi-modal music emotion. In: **11th International Society for Music Information and Retrieval Conference**. [S.l.: s.n.], 2010. p. 105–105.

MALHEIRO, R.; PANDA, R.; GOMES, P.; PAIVA, R. P. Emotionally-Relevant Features for Classification and Regression of Music Lyrics. **IEEE Transactions on Affective Computing**, v. 9, n. 2, p. 240–254, abr. 2018. ISSN 1949-3045.

MCFEE, B.; RAFFEL, C.; LIANG, D.; ELLIS, D. P.; MCVICAR, M.; BATTENBERG, E.; NIETO, O. librosa: Audio and music signal analysis in python. In: . [S.l.: s.n.], 2015.

MCKAY, C.; FUJINAGA, I.; DEPALLE, P. jaudio: A feature extraction library. In: **Proceedings of the International Conference on Music Information Retrieval**. [S.l.: s.n.], 2005. p. 600–603.

MCVICAR, M.; FREEMAN, T.; BIE, T. D. Mining the correlation between lyrical and audio features and the emergence of mood. In: **International Society for Music Information Retrieval**. [S.l.: s.n.], 2011. p. 783–788.

MERMELSTEIN, P. Distance measures for speech recognition, psychological and instrumental. **Pattern recognition and artificial intelligence**, v. 116, p. 374–388, 1976.

MIHALCEA, R.; STRAPPARAVA, C. Lyrics, Music, and Emotions. In: **Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning**. Jeju Island, Korea: Association for Computational Linguistics, 2012. p. 590–599. Disponível em: <http://www.aclweb.org/anthology/D12-1054>.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

OJALA, T.; PIETIKAINEN, M.; MAENPAA, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. **IEEE Transactions on pattern analysis and machine intelligence**, IEEE, v. 24, n. 7, p. 971–987, 2002.

PANDA, R.; MALHEIRO, R.; PAIVA, R. P. Musical texture and expressivity features for music emotion recognition. In: **International Society for Music Information Retrieval**. [S.l.: s.n.], 2018. p. 383–391.

PATRA, B. G.; DAS, D.; BANDYOPADHYAY, S. Mood Classification of Hindi Songs based on Lyrics. In: **Proceedings of the 12th International Conference on Natural Language Processing**. Trivandrum, India: NLP Association of India, 2015. p. 261–267. Disponível em: <http://www.aclweb.org/anthology/W15-5939>.

PATRA, B. G.; DAS, D.; BANDYOPADHYAY, S. Multimodal mood classification of hindi and western songs. **Journal of Intelligent Information Systems**, Springer, p. 1–18, 2018.

PAULINO, M. A. D.; COSTA, Y. M. e Gomes da; JUNIOR, A. S. B.; SVAIGEN, A. R.; AYLON, L. B. R.; OLIVEIRA, L. E. S. de. A brazilian speech database. **2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)**, p. 234–241, 2018.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PRZYBYSZ, A. L. **Classificação Automática de Emoções em Músicas Latinas Utilizando Diferentes Fontes de Informação**. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2016.

QUINLAN, J. R. Simplifying decision trees. **International journal of man-machine studies**, Elsevier, v. 27, n. 3, p. 221–234, 1987.

REHUREK, R.; SOJKA, P. Gensim–python framework for vector space modelling. **NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic**, v. 3, n. 2, 2011.

RIBEIRO, R. do P. **Classificação Automática de Emoções em Letras de Música Utilizando Combinação de Classificadores**. 2015. Monografia (Bacharel em Engenharia de Computação), Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Brazil.

RIBEIRO, R. P.; ALMEIDA, M. A.; JR, C. N. S. The ethnic lyrics fetcher tool. **EURASIP Journal on Audio, Speech, and Music Processing**, Nature Publishing Group, v. 2014, n. 1, p. 27, 2014.

RUSSELL, J. A. A circumplex model of affect. **Journal of personality and social psychology**, American Psychological Association, v. 39, n. 6, p. 1161, 1980.

SANTOS, C. L. D.; SILLA, C. N. The latin music mood database. **EURASIP Journal on Audio, Speech, and Music Processing**, Nature Publishing Group, v. 2015, n. 1, p. 23, 2015.

SCHAPIRE, R. E.; FREUND, Y. Boosting: Foundations and algorithms. Emerald Group Publishing Limited, 2012.

SCHULLER, B.; WENINGER, F.; DORFNER, J. Multi-modal non-prototypical music mood analysis in continuous space: Realiability and performances. In: **Proc. 12th International Society for Music Information Retrieval 2011, ISMIR, Miami, FL, USA**. [S.l.: s.n.], 2011.

SILLA, C. N.; KOERICH, A. L.; KAESTNER, C. A. The latin music database. In: **International Society for Music Information Retrieval**. [S.l.: s.n.], 2008. p. 451–456.

SOUZA, V. M. A. D. **Análise automática de coerência semântica em resumos acadêmicos escritos em português**. Dissertação (Mestrado) — Universidade Estadual de Maringá, 2011.

SOUZA, V. M. A. de. **Análise automática de coerência semântica em resumos acadêmicos escritos em português**. Dissertação (Mestrado) — Universidade Estadual de Maringá, 2011.

STRAPPARAVA, C.; MIHALCEA, R.; BATTOCCHI, A. A Parallel Corpus of Music and Lyrics Annotated with Emotions. In: **Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)**. Istanbul, Turkey: European Language Resources Association (ELRA), 2012. Disponível em: <http://www.aclweb.org/anthology/L12-1425>.

SU, D.; FUNG, P. These words are music to my ears: Recognizing music emotion from lyrics using AdaBoost. In: **2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference**. [S.l.: s.n.], 2013. p. 1–4.

SU, D.; FUNG, P.; AUGUIN, N. Multimodal music emotion classification using AdaBoost with decision stumps. In: **2013 IEEE International Conference on Acoustics, Speech and Signal Processing**. [S.l.: s.n.], 2013. p. 3447–3451.

TAN, K.; VILLARINO, M.; MADERAZO, C. Automatic music mood recognition using russell's twodimensional valence-arousal space from audio and lyrical data as classified using svm and naïve bayes. In: IOP PUBLISHING. **IOP Conference Series: Materials Science and Engineering**. [S.l.], 2019. v. 482, n. 1, p. 012019.

TAVARES, J. C. C.; MALDONADO, Y.; COSTA, G. da. Music mood classification using visual and acoustic features. In: IEEE. **2017 XLIII Latin American Computer Conference (CLEI)**. [S.l.], 2017. p. 1–10.

THAYER, R. E. **The biopsychology of mood and arousal**. [S.l.]: Oxford University Press, 1990.

TZANETAKIS, G.; COOK, P. Marsyas A framework for audio analysis. **Organised sound**, Cambridge University Press, v. 4, n. 3, p. 169–175, 2000.

TZANETAKIS, G.; COOK, P. Musical genre classification of audio signals. **IEEE Transactions on speech and audio processing**, IEEE, v. 10, n. 5, p. 293–302, 2002.

UJLAMBKAR, A. M.; ATTAR, V. Z. Mood Classification of Indian Popular Music. In: **Proceedings of the CUBE International Information Technology Conference**. New York, NY, USA: ACM, 2012. (CUBE '12), p. 278–283. ISBN 978-1-4503-1185-4. Disponível em: <http://doi.acm.org/10.1145/2381716.2381768>.

VALE, P. M. F. **The role of artist and genre on music emotion recognition**. Tese (Doutorado), 2017.

VAPNIK, V. N. The nature of statistical learning. **Theory**, springer, 1995.

WANG, X.; CHEN, X.; YANG, D.; WU, Y. Music emotion classification of chinese songs based on lyrics using tf* idf and rhyme. In: CITESEER. **International Society for Music Information Retrieval**. [S.l.], 2011. p. 765–770.

WANG, X.; WU, Y.; CHEN, X.; YANG, D. Enhance popular music emotion regression by importing structure information. In: **2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference**. [S.l.: s.n.], 2013. p. 1–4.

WANG, Z.; ZHANG, J.; VERMA, N. Realizing low-energy classification systems by implementing matrix multiplication directly within an adc. **IEEE transactions on biomedical circuits and systems**, IEEE, v. 9, n. 6, p. 825–837, 2015.

WATSON, D.; MANDRYK, R. L. Modeling musical mood from audio features and listening context on an in-situ data set. 2012.

WATSON, D.; TELLEGEN, A. Toward a consensual structure of mood. **Psychological bulletin**, US: American Psychological Association, v. 98, n. 2, p. 219, 1985.

WITTEN, I. H.; FRANK, E.; HALL, M. A. Practical machine learning tools and techniques. **Morgan Kaufmann**, p. 578, 2005.

WU, B.; ZHONG, E.; HORNER, A.; YANG, Q. Music Emotion Recognition by Multi-label Multi-layer Multi-instance Multi-view Learning. In: **Proceedings of the 22Nd ACM International Conference on Multimedia**. New York, NY, USA: ACM, 2014. (MM '14), p. 117–126. ISBN 978-1-4503-3063-3. Disponível em: <http://doi.acm.org/10.1145/2647868.2654904>.

XIA, Y.; WANG, L.; WONG, K.-F.; XU, M. Lyric-based Song Sentiment Classification with Sentiment Vector Space Model. In: **Proceedings of ACL-08: HLT, Short Papers**. Columbus, Ohio: Association for Computational Linguistics, 2008. p. 133–136. Disponível em: <http://www.aclweb.org/anthology/P08-2034>.

XIONG, Y.; SU, F.; WANG, Q. Automatic music mood classification by learning cross-media relevance between audio and lyrics. In: **2017 IEEE International Conference on Multimedia and Expo (ICME)**. [S.l.: s.n.], 2017. p. 961–966.

YANG, D.; LEE, W. Music Emotion Identification from Lyrics. In: **2009 11th IEEE International Symposium on Multimedia**. [S.l.: s.n.], 2009. p. 624–629.

ZAANEN, M. V.; KANTERS, P. Automatic mood classification using tf* idf based on lyrics. In: **International Society for Music Information Retrieval**. [S.l.: s.n.], 2010. p. 75–80.

ZHANG, S.; REPETTO, R. C.; SERRA, X. Understanding the expressive functions of jingju metrical patterns through lyrics text mining. 2017.

ZHOU, Z.-H. Ensemble learning. **Encyclopedia of biometrics**, v. 1, p. 270–273, 2009.

APÊNDICE A – PROTOCOLO DA REVISÃO SISTEMÁTICA

A seguir é apresentado o protocolo utilizado para a realização da revisão sistemática apresentada na Seção 3.1.

A.1 PLANEJAMENTO DA REVISÃO SISTEMÁTICA

O objetivo da revisão sistemática realizada neste trabalho foi identificar e interpretar trabalhos relacionados à tarefa de classificação de músicas por emoções baseando-se nas letras das músicas. Os trabalhos abordados apontaram três fatores importantes para alcançar o objetivo, a saber: estudo das emoções e seus modelos ou taxonomias de representação; os métodos utilizados para a classificação de músicas; e as características extraídas das letras para serem utilizadas na classificação.

A.1.1 QUESTÕES DA PESQUISA

Baseando-se nos objetivos descritos anteriormente, as seguintes questões foram propostas:

- **Questão 1**: Quais modelos ou taxonomias de emoções são utilizadas para categorizar as músicas?
- **Questão 2**: Quais características podem ser extraídas das letras de músicas para serem utilizadas na classificação de músicas por emoções?
- **Questão 3**: Quais algoritmos de aprendizado estão sendo utilizados na área de MER?

A resposta da primeira questão permite compreender como as emoções podem ser categorizadas e assim entender padrões ou comportamentos. A partir da segunda questão é possível realizar um levantamento detalhado das características que podem ser utilizadas no processo de classificação e assim combina-las por meio de experimentos realizados com

os algoritmos encontrados na literatura. Por fim, respondendo a terceira questão é possível observar os algoritmos propostos e suas formas de aplicação no contexto da recuperação de informações musicais.

A.1.2 *STRING* DE BUSCA

Após a definição das questões da pesquisa, foi criada a *string* de busca baseada nos termos mais frequentes relacionados à tarefa de classificação de músicas por emoções baseando-se nas letras das músicas. Foram incluídos na revisão sistemática trabalhos que exploraram unicamente letras e trabalhos que utilizaram letras juntamente com outras fontes de informação. A *string* de busca foi aplicada nas seguintes bases de artigos científicos:

- IEEE (Institute of Electrical and Electronics Engineers) `http://ieeexplore.ieee.org/`;
- ACM (*Association for Computing Machinery*) `https://dl.acm.org/`;
- Science Direct `https://www.sciencedirect.com/`;
- Springer `https://link.springer.com/`;
- ACL (Association for Computational Linguistics) http://aclweb.org/anthology/;

A *string* de busca utilizada foi a seguinte:

- (music *OR* song) *AND* (mood *OR* emotion) *AND* (classification *OR* identification *OR* recognition) *AND* lyrics.

A.1.3 CRITÉRIOS DE INCLUSÃO E EXCLUSÃO

A terceira fase do planejamento da revisão sistemática é estabelecer os critérios de inclusão e exclusão que são utilizados para filtrar os artigos recuperados. Os critérios de inclusão e exclusão usados neste trabalho são apresentados, respectivamente, nos Quadros 10 e 11.

Quadro 10: Critérios de Inclusão

Critério	Descrição
C1	Pub. que propõem classificadores de músicas por emoções.
C2	Pub. que utilizam a letra como fonte de informação.

Fonte: Autoria Própria (2018)

Quadro 11: Critérios de Exclusão

Critério	Descrição
C3	Pub. com acesso restrito, que não estão em inglês e ou que contenham menos de 4 páginas.
C4	Pub. que são posters, apresentações ou tutoriais.
C5	Pub. que são secundárias (surveys).

Fonte: Autoria Própria (2018)

A.2 CONDUÇÃO

Após a fase de planejamento ser concluída, a revisão sistemática foi aplicada. Nessa etapa, o primeiro passo foi aplicar a *string* de busca nas bases selecionadas. Foram encontrados 89 trabalhos. Na Tabela 36 são apresentadas as quantidades de artigos encontrados em cada base selecionada.

Tabela 36: Número de trabalhos científicos encontrados em cada base pesquisada

Bases	Total de artigos
IEEE	28
ACM	10
Science Direct	1
Springer	42
ACL	8
Total	**89**

Fonte: Autoria Própria (2018)

Após a pesquisa com a *string* de busca, todos os 89 trabalhos encontrados passaram pelos critérios de inclusão e exclusão. Por meio da aplicação desses critérios foram aprovados 25 artigos. O critério C1 não incluiu 49 trabalhos e o critério C2 não incluiu 49 trabalhos, sendo

que existe uma interseção entre a quantidade de trabalhos não incluídos por esses dois critérios. O critério C3 excluiu dois trabalhos. Por fim, os critérios C4 e C5 excluíram três artigos cada.

Foram adicionados manualmente 10 artigos da ISMIR (*International Society of Music Information Retrieval*). Esses artigos não estavam indexados nas bases utilizadas, por isso foram adicionados manualmente. Assim, cada um dos 25 artigos encontrados nas bases e os 10 artigos adicionados manualmente foram analisados e deles foram extraímos os dados conforme apresentado na Subseção A.2.1.

A.2.1 EXTRAÇÃO DE DADOS

Por fim, foram extraídos os seguintes dados dos artigos aprovados:

- Dados estatísticos:
  - Ano de publicação dos trabalhos;
  - Bases científicas utilizadas.
- Bases de dados:
  - Fontes das letras das músicas;
  - Quantidade de músicas na base;
  - Língua das letras da base;
  - Gêneros musicais.
- Modelos de emoções e taxonomias:
  - Modelos de emoções e taxonomias utilizados como base para a anotação das músicas.
- Características e Algoritmos de aprendizado utilizados:
  - Fontes de informação;
  - Características extraídas das letras;
  - Tipos de classificação (monorótulo ou multirótulo).

Com esses dados é possível responder as questões de pesquisa formuladas. Além disso, foram extraídas informações extras que podem auxiliar em trabalhos futuros.

APÊNDICE B – EXPERIMENTAÇÃO DE PARÂMETROS PARA D2V

A seguir são apresentadas as configurações utilizadas e os resultados dos experimentos realizados para definir os parâmetros a serem empregados nos experimentos relacionados ao D2V. Na tabela 37, os nomes dos parâmetros foram abreviados da seguinte forma: *Vector size* (V), *epochs* (e), *window* (w) e *dbow_words* (DbW). A configuração escolhida foi *V=250, e=100, w=1*.

Tabela 37: Resultado da experimentação de parâmetros do D2V

Configurações	Alegria	Amor	Decep.	Exc./Ent.	Paixão	Tris.	Média
V=250, e=100	0,361	0,472	0,068	0,547	0,387	0,219	**0,396**
V=100, e=100	0,404	0,489	0,091	0,571	0,388	0,268	**0,418**
V=250, e=20	0,395	0,476	0,000	0,572	0,409	0,234	**0,411**
V=100, e=20	0,375	0,498	0,000	0,567	0,403	0,190	**0,406**
V=250, e=20, w=100	0,171	0,405	0,000	0,231	0,331	0,117	**0,271**
V=100, e=20, w=1	0,387	0,491	0,066	0,606	0,373	0,236	**0,412**
V=250, e=20, w=1	0,348	0,482	0,000	0,614	0,400	0,215	**0,407**
V=100, e=100, w=1	0,398	0,477	0,000	0,592	0,374	0,239	**0,406**
V=250, e=100, w=1	0,419	0,473	0,124	0,617	0,392	0,300	**0,430**
V=250, e=100, w=1, alpha=0.5	0,021	0,296	0,000	0,021	0,354	0,023	**0,180**
V=250, e=100, w=1, alpha=0.01	0,402	0,484	0,038	0,624	0,420	0,259	**0,430**
V=250, e=100, w=1, alpha=0.025, hs=1	0,270	0,408	0,017	0,481	0,332	0,180	**0,334**
V=250, e=100, w=1, alpha=0.025, DbW=1	0,270	0,408	0,017	0,481	0,332	0,180	**0,334**

Fonte: Autoria Própria (2020)

APÊNDICE C – EXPERIMENTAÇÃO DE PARÂMETROS PARA LSA

A seguir são apresentados os resultados dos experimentos realizados para definir a quantidade de tópicos (dimensão) a ser utilizada nos demais experimentos relacionados a LSA. Como pode ser observado na Tabela 38, a quantidade de 100 tópicos obteve a melhor média e, portanto, foi escolhida para ser utilizada nos experimentos.

Tabela 38: Resultado da experimentação de parâmetros LSA

Tópicos	Alegria	Amor	Decep.	Exc./Ent.	Paixão	Tris.	Média
10	0,360	0,500	0,000	0,577	0,388	0,000	**0,375**
30	0,405	0,499	0,025	0,594	0,402	0,112	**0,404**
50	0,405	0,490	0,048	0,585	0,399	0,157	**0,407**
95	0,403	0,471	0,022	0,589	0,388	0,180	**0,401**
97	0,404	0,459	0,043	0,595	0,396	0,178	**0,402**
100	0,407	0,473	0,064	0,601	0,398	0,204	**0,412**
103	0,403	0,471	0,080	0,598	0,382	0,185	**0,404**
105	0,387	0,471	0,063	0,597	0,386	0,189	**0,402**
250	0,395	0,447	0,113	0,548	0,348	0,230	**0,387**
500	0,333	0,374	0,092	0,510	0,296	0,200	**0,334**
1.000	0,302	0,386	0,072	0,475	0,297	0,221	**0,330**
2.237	0,275	0,361	0,058	0,432	0,293	0,238	**0,314**

Fonte: Autoria Própria (2020)

APÊNDICE D – QUANTIDADE DE CARACTERÍSTICAS EXTRAÍDAS

A seguir são apresentados os tamanhos dos vetores de características empregados nos diferentes experimentos realizados neste trabalho. Nas Tabelas 39 e 40 são apresentados os tamanhos dos vetores de características extraídos das letras, bem como da combinação entre elas. De forma análoga, na Tabela 41 são apresentados os tamanhos dos vetores das características extraídos dos áudios. Por fim, nas Tabelas 42-45 são apresentados os tamanhos dos vetores resultantes das combinações de características entre letras e áudios.

Tabela 39: Dimensões dos vetores de características baseados em TF-IDF e características estilísticas.

Config.	Qtde.	Config.	Qtde.
2G	614	ST+EST	13.814
3G	4.590	2G+ST	14.205
4G	17.171	3G+ST	17.251
ST	13.798	4G+ST	28.649
EST	16	2G+ST+EST	12.421
2G+EST	630	3G+ST+EST	17.267
3G+EST	4.606	4G+ST+EST	28.665
4G+EST	17.187		

Fonte: Autoria Própria (2020)

Tabela 40: Quantidade de características por experimentos baseados em LSA, D2V e características estilísticas

Config.	Qtde.	Config.	Qtde.
2G	100	2G+ST	200
3G	100	3G+ST	200
4G	100	4G+ST	200
ST	100	2G+ST+EST	216
EST	16	3G+ST+EST	216
2G+EST	116	4G+ST+EST	216
3G+EST	116	D2V	250
4G+EST	116	D2V+EST	266
ST+EST	116		

Fonte: Autoria Própria (2020)

Tabela 41: Quantidade de características por experimentos baseados em características texturais e acústicas

Config.	Qtde.	Config.	Qtde.
RLBP	59	SSD+RH+RP	1.668
MFCC-med	20	RLBP+TIMB.	107
MFCC-med+var	40	RLBP+SSD	1.727
TIMBRAL-med+var	48	TIMB.+SSD	1.716
somente SSD	168	RLBP+TIMB.SSD	1.775

Fonte: Autoria Própria (2020)

Tabela 42: Quantidade de características por experimentos multimodais para classificação de emoção

Config.	Qtde.	Config.	Qtde.
3G+EST+RLBP	4.665	3G(lsa)+ST(lsa)+RLBP+TIMB.+SSD	1.975
3G+EST+RLBP+TIMB.	4.713	D2V+RLBP	309
3G+EST+RLBP+TIMB.+SSD	6.381	D2V+TIMB.	298
3G(lsa)+ST(lsa)+RLBP	259	D2V+SSD	1.918
3G(lsa)+ST(lsa)+RLBP+TIMB.	307		

Fonte: Autoria Própria (2020)

Tabela 43: Quantidade de características por experimentos multimodais para classificação de valência

Config.	Qtde.	Config.	Qtde.
2G+ST+RLBP	14.264	3G(lsa)+ST(lsa)+RLBP+TIMB.+SSD	1.975
2G+ST+TIMB.	14.253	D2V+EST+RLBP	325
2G+ST+SSD	15.873	D2V+EST+RLBP+TIMB.	373
3G(lsa)+ST(lsa)+RLBP	259	D2V+EST+RLBP+SSD	1.993
3G(lsa)+ST(lsa)+RLBP+TIMB.	307		

Fonte: Autoria Própria (2020)

Tabela 44: Quantidade de características por experimentos multimodais para classificação de excitação

Config.	Qtde.	Config.	Qtde.
4G+RLBP	17.230	3G(lsa)+SSD	1.768
4G+TIMB.	17.219	D2V+RLBP	309
4G+SSD	18.839	D2V+TIMB.	298
3G(lsa)+RLBP	159	D2V+SSD	1.918
3G(lsa)+TIMB.	148		

Fonte: Autoria Própria (2020)

Tabela 45: Quantidade de características por experimentos multimodais para classificação de quadrante

Config.	Qtde.	Config.	Qtde.
3G+RLBP	4.649	3G(lsa)+ST(lsa)+EST +RLBP+TIMB.+SSD	1.991
3G+TIMB.	4.638	D2V+RLBP	309
3G+SSD	6.258	D2V+RLBP+TIMB.	357
3G(lsa)+ST(lsa)+EST+RLBP	275	D2V+RLBP+SSD	1.977
3G(lsa)+ST(lsa)+EST+RLBP+TIMB.	323		

Fonte: Autoria Própria (2020)