Universidade de São Paulo
Escola Politécnica
Programa de Educação Continuada em Engenharia
Especialização em Inteligência Artificial

Lucas Cabral Gomes

Identificando músicas com potencial de sucesso no gênero do rap

São Paulo
2024

LUCAS CABRAL GOMES

Identificando músicas com potencial de sucesso no gênero do rap

— Versão Original —

Monografia apresentada ao Programa de Educação Continuada em Engenharia da Escola Politécnica da Universidade de São Paulo como parte dos requisitos para conclusão do curso de Especialização em Inteligência Artificial.

Orientador: Prof. Dr. Marcos Lopes

São Paulo
2024

Catalogação-na-publicação

Gomes, Lucas Cabral
Identificando músicas com potencial de sucesso no gênero do rap/ L.Gomes – São Paulo, 2024.
150p.

Monografia (Especialização em Inteligência Artificial) – Escola Politécnica da Universidade de São Paulo. PECE – Programa de Educação Continuada em Engenharia.

1. Processamento de linguagem natural 2. Análise musical 3. Rap 4. Word2Vec 5. TF-IDF 6. Floresta Aleatória 7. Regressão Logística.
I. Universidade de São Paulo. Escola Politécnica. PECE – Programa de Educação Continuada em Engenharia. II.t.

A minha sorte.

Agradecimentos

Agradeço primeiramente aos meus pais, *Luiz* e *Cláudia*, que sempre me incentivaram a buscar conhecimento e investiram em minha educação.

À minha namorada *Thays*, por me apoiar incondicionalmente e por dividir sonhos comigo.

Aos meus colegas de trabalho *Alexandre Coutinho* e *Marco Barbosa*, por acreditarem em mim e por me apresentarem grandes desafios.

Aos professores do Programa de Educação Continuada da Poli-USP, em especial à Profa. Dra. *Larissa Driemeier* e ao Prof. Dr. *Thiago de Castro Martins* por coordenarem o curso com tanta paixão e dedicação, e ao Prof. Dr. *Marcos Lopes* pela parceria, disponibilidade e pelos valiosos ensinamentos.

E, por fim, aos meus colegas de turma, pelos estudos que compartilhamos juntos.

Boa sorte é o que acontece quando a
oportunidade encontra o planejamento.

— *Thomas Edison*

Sumário

Sumário • *v*

Resumo • *vii*

Abstract • *viii*

Lista de Figuras • *ix*

Lista de Tabelas • *x*

1 Introdução • *1*

2 Revisão da literatura • *4*
2.1 O movimento *hip hop* e o *rap* • 4
2.2 *Hit Song Science* • 4
2.3 *Word embeddings* • 5
2.4 Redução de dimensionalidade com PCA • 6
2.5 Análise textual de canções • 6
2.6 Aprendizado de Máquina e classificação binária • 7
2.7 Transformers • 8

3 Metodologia • *9*
3.1 Coleta de Dados • 9
3.2 Seleção da variável-alvo • 10
3.3 Tratamento de Dados • 11
3.3.1 *Tokenização* • 11
3.3.2 Representação por vetores numéricos • 12
3.3.3 Redução de Dimensionalidade • 13
3.3.4 Métricas Auxiliares • 13
3.4 Conjuntos de Dados • 15
3.5 Seleção de Modelos • 16

4 Resultados e Discussão • *18*
4.1 Análise por representações vetoriais • 18
4.1.1 Visualização dos dados • 18
4.1.2 Categorização • 21

5 Conclusão • *26*

5.1 Limitações e Trabalhos Futuros • 26

Referências • *28*

Resumo

GOMES, L. *Identificando músicas com potencial de sucesso no gênero do rap.* 2024. Monografia (Especialização em Inteligência Artificial) – Escola Politécnica da Universidade de São Paulo. PECE – Programa de Educação Continuada em Engenharia. Universidade de São Paulo, São Paulo, 2024.

O objetivo desta pesquisa foi identificar de maneira consistente canções do gênero de *rap* com potencial comercial, sendo identificadas por "Mais ouvidas" ou "Menos ouvidas", a partir de suas transcrições. Para essa tarefa, foram aplicadas técnicas de processamento de linguagem natural (PLN) para a limpeza e transformações dos dados, assim como técnicas de vetores de palavras, como TF-IDF e Word2Vec, para a criação de representações numéricas das transcrições das letras das canções capazes de capturar o contexto e importância de cada termo empregado nas músicas. Também são calculadas métricas auxiliares, como indicadores de diversidade lexical, representatividade de classes gramaticais, uso de gírias e xingamentos, análise de sentimento e de subjetividade, entre outros, complementares às representações de vetores numéricos de palavras. Por fim, foram criados classificadores binários baseados em cinco modelos de aprendizado de máquina com características distintas, e foi realizada uma busca exaustiva pelos melhores hiperparâmetros de cada modelo com o objetivo de se obter os melhores resultados. Os resultados identificam o modelo de Floresta Aleatória combinado com as representações numéricas geradas pelo método TF-IDF como o mais bem avaliado na tarefa, tendo obtido tanto acurácia quanto Medida-F de 81,8%. Por outro lado, o modelo de regressão logística foi o mais bem avaliado ao utilizar as métricas auxiliares, tendo obtido acurácia e Medida-F de 77,5%. As duas técnicas combinadas estabelecem um método consistente para a identificação de músicas com bom potencial comercial. Por contar com um corpus de apenas 2.000 músicas, a pesquisa se limita a modelos de aprendizado de máquina menos complexos, sendo possível aplicar modelos mais complexos (isto é, de aprendizado profundo) com um corpus maior. Outra investigação interessante para o futuro é o uso de grandes modelos de linguagem pré-treinados e ajustados para essa tarefa, com vistas à obtenção de melhores resultados.

Palavras-chave: Processamento De Linguagem Natural. Análise Musical. Rap. Word2Vec. TF-IDF. Floresta Aleatória. Regressão Logística

Abstract

GOMES, L.*Identifying songs with potential for success in the rap genre.* 2024. Monografia (Especialização em Inteligência Artificial) – Escola Politécnica da Universidade de São Paulo. PECE – Programa de Educação Continuada em Engenharia. University of São Paulo, São Paulo, Brazil. 2024.

The aim of this research was to consistently identify songs in the *rap* genre with commercial potential, identified by "Most listened" or "Least listened" from their transcriptions. For this task, natural language processing techniques were applied to clean and transform the data, as well as *word embeddings* techniques such as TF-IDF and *Word2Vec* to create numerical representations of the song transcriptions capable of capturing the context and importance of each term used in the songs. Auxiliary metrics are also calculated, such as indicators of lexical diversity, representativeness of grammatical classes, use of slang and swear words, analysis of sentiment and subjectivity, among others, complementary to the representations of numerical vectors. Finally, five models with different characteristics were used, and an exhaustive search was carried out for the best hyperparameters for each model in order to obtain the best results. At the end of the research, it was found that the Random Forest model combined with the numerical representations generated by the TF-IDF method was the best evaluated in the task, having obtained an accuracy and F-Measure of 81,8%. On the other hand, the logistic regression model was the best evaluated when using the auxiliary metrics, obtaining an accuracy and F-Measure of 77,5%. The two techniques combined establish a consistent method for identifying songs with commercial potential. As the corpus is only 2000 songs, the research is limited to less complex models. It is possible to apply more complex models with a larger corpus or even to use pre-trained models with refinement to obtain better results.

Keywords: Natural Language Processing. Musical Analysis. Rap. Word2Vec. TF-IDF. Random Forest. Logistic Regression

Lista de Figuras

2.1 Matriz de confusão utilizada para calcular as métricas de avaliação descritas, como acurácia, precisão, cobertura e medida-f. 7

3.1 Diagrama dos processos descritos na metodologia. 9

4.1 Distribuição de frequência de palavras após a etapa de limpeza e pré-processamento dos dados. 19
4.2 Representação 2D dos componentes extraídos a partir da representação vetorial gerada pelo algorítimo Word2Vec. 19
4.3 Representação 3D dos componentes extraídos a partir da representação vetorial gerada pelo algorítimo Word2Vec. 20
4.4 Representação 2D dos componentes extraídos a partir da representação vetorial gerada pelo algorítimo TF-IDF. 20
4.5 Representação 3D dos componentes extraídos a partir da representação vetorial gerada pelo algorítimo TF-IDF. 20

Lista de Tabelas

4.1 Métricas de Desempenho dos modelos utilizando Word2Vec. 21
4.2 Métricas de Desempenho dos modelos utilizando TF-IDF. 21
4.3 Métricas de Desempenho dos modelos utilizando os N primeiros componentes gerados a partir do TF-IDF que explicam 80% da variância dos dados). 22
4.4 Métricas de Desempenho dos modelos utilizando as métricas auxiliares. . . . 23
4.5 Comparativo do modelo de Regressão Logística antes e após o processo de seleção de variáveis. 23
4.6 Lista de variáveis restantes após o processo de seleção de variáveis e seus coeficientes. 23
4.7 Métricas de Desempenho do Modelo Ensemble utilizando características combinadas de TF-IDF, Word2Vec e Métricas Auxiliares. 25

Capítulo 1

Introdução

A indústria musical, sendo o quinto maior segmento do mercado de entretenimento, faturou cerca de 11.4 bilhões de dólares em 2022, refletindo um crescimento aproximado de 28% em relação ao ano anterior. Comparada aos quatro primeiros mercados de entretenimento (Video Games, Televisão aberta, *Smart TV* e Cinema), a industria musical apresenta muito menos barreiras de entrada para artistas independentes e pequenos produtores, que podem produzir suas músicas com um baixo investimento e divulgá-las em plataformas de *streaming* e redes sociais de forma gratuita, representando muitas vezes um sonho de melhor qualidade de vida e independência financeira obtida através da arte para jovens das mais diversas classes sociais.

A cultura *hip hop*, que teve início nos Estados Unidos na década de 1970 em comunidades negras, tem como representante do aspecto musical o gênero *rap*, que possui uma fala rítmica e rimada, e que dá voz à jovens da periferia para que possam expressar suas dificuldades do dia a dia e suas aspirações. Através do *rap*, muitos jovens fizeram sucesso tanto no aspecto financeiro quanto na tentativa de trazer luz para temas sensíveis presentes na periferia como violência, repressão policial, drogas, miséria e crime, e foram capazes de melhorar a qualidade de vida em suas comunidades assim como a incentivar outros jovens a seguirem a mesma trajetória. Exemplos bem conhecidos são Jay-Z, Nas, Tupac, Eminem e Snoop Dog, entre outros artistas.

Dada a importância cultural do gênero *rap* e o impacto financeiro positivo que pode trazer tanto para os artistas como para todos os profissionais envolvidos, a proposta desta pesquisa é de encontrar as características mais relevantes na identificação de canções de sucesso, ou seja, de grande audiência, para possibilitar tanto a identificação de novos artistas promissores como, também, a orientação quanto às características comuns dos produtos musicais de maior sucesso.

Para a realização desta pesquisa, foram coletadas as transcrições de duas mil canções do gênero do *rap*, categorizadas de forma balanceada entre “Mais ouvidas”e “Menos

ouvidas", a partir do número de reproduções na plataforma Last.fm[1]. Sobre esses dados foram aplicadas técnicas de Processamento de Linguagem Natural (PLN) em uma etapa de pré-processamento, onde tokenizaram-se as transcrições das canções visando eliminar as *stop words*, converter palavras com contrações, padronizar termos com grafias diferentes, mas significados similares, e extrair lemas.

Após o pré-processamento, foram aplicadas técnicas de vetorização de palavras para criar representações numéricas das transcrições, capazes de extrair contexto e a importância dos termos empregados em cada música.

Também foram calculadas métricas auxiliares, como diversidade lexical, uso de gírias e xingamentos, representatividade por classe gramatical, análise de subjetividade e sentimento, entre outras, com o objetivo de complementar as representações por vetores numéricos.

Por fim, foram testados diversos modelos probabilísticos de aprendizado de máquina, com diferentes combinações de conjuntos de dados, e foi empregada a técnica de *grid search*, onde se realiza uma busca exaustiva de hiperparâmetros para os modelos a fim de se obter melhores resultados calculados a partir de métricas de avaliação.

O melhor modelo obtido foi o de Floresta Aleatória, com refinamento de hiperparâmetros e com dados de entrada gerados pelo método TF-IDF, que conseguiu identificar entre canções "Mais ouvidas"e "Menos ouvidas"com 81,8% de acurácia e de Medida-F. Por outro lado, o modelo de regressão logística foi o melhor na identificação entre as duas categorias utilizando as métricas auxiliares, após um processo de seleção de variáveis e refinamento de hiperparâmetros, obtendo acurácia e Medida-F iguais a 77,5%.

Também foram identificadas quais características são mais relevantes na identificação de canções com potencial comercial, sendo a característica mais importante a utilização de linguagem emocionalmente expressiva e de teor pessoal.

Dessa forma, combinar a análise das representações numéricas geradas por técnicas de vetores de palavras com o uso de métricas auxiliares revela-se um método consistente para identificar canções do gênero de *rap* com potencial comercial.

O leitor encontrará no Capítulo 2 a revisão da literatura, onde são apresentadas técnicas de processamento de linguagem natural empregadas em trabalhos de temática semelhante, assim como o contexto histórico do gênero do *rap*. No Capítulo 3, são discutidas detalhadamente as técnicas aplicadas nesta pesquisa, com início no processo de coleta de dados, a limpeza e transformação dos dados, a criação de representações numéricas, cálculos de métricas auxiliares e o processo de seleção de modelos. No Capítulo 4 são apresentados os resultados deste estudo e é feita a comparação entre as técnicas praticadas, identificando os principais fatores na distinção de canções com

[1] https://www.last.fm/

potencial comercial. O Capítulo 5 fecha este trabalho com uma breve contextualização dos resultados frente à literatura existente, assim como indica as limitações e possíveis encaminhamentos futuros desta pesquisa.

Capítulo 2

Revisão da literatura

Neste capítulo será apresentada a revisão bibliográfica referente à origem do *rap* e sua influência cultural, o desenvolvimento da *Hit Song Science*, o processamento de linguagem natural aplicada à análise musical, incluindo algoritmos de aprendizado de máquina utilizados para classificação binária e o uso de redes *Transformers* nessas tarefas.

2.1 O movimento *hip hop* e o *rap*

A cultura *hip hop*, que teve início nos Estados Unidos na década de 1970 (Lourenço 2010), em comunidades negras, tem o *rap* como representante do aspecto musical, que possui uma fala rítmica e rimada e dá voz a jovens da periferia, permitindo-lhes expressar suas dificuldades cotidianas e aspirações.

Através do *rap*, muitos jovens alcançaram sucesso não só financeiro, mas também na exposição de temas sensíveis da periferia, como violência, repressão policial, drogas, miséria e crime, conseguindo melhorar a qualidade de vida em suas comunidades (Miranda 2016) e incentivando outros jovens a seguir o mesmo caminho. Exemplos relevantes incluem Jay-Z, Nas, Tupac, Eminem, Snoop Dog entre outros.

2.2 *Hit Song Science*

Hit Song Science é um termo registrado pela Polyphonic HMI, empresa pioneira em recuperação de informações musicais (MIR), que originou um produto homônimo, que consiste num algoritmo com finalidade de identificar quais músicas possuem maior potencial de sucesso comercial (Herremans 2019).

O algoritmo ganhou reconhecimento ao identificar músicas com grande potencial, exemplificado pelo álbum de 2002 da cantora Norah Jones, "Come Away With Me", que vendeu mais de 23 milhões de cópias, desafiando opiniões céticas. Outro exemplo notável

foi a música "Hey Ya!" do grupo OutKast, lançada em 2003, que alcançou o topo das paradas, embora isso tenha requerido também um condicionamento do público por meio de ações de marketing, conforme citado no livro (Duhigg 2012).

O sucesso do algoritmo não se limita ao início dos anos 2000, como observado pela revista VICE em (Neal 2015), quando, em 2015, o algoritmo previu com mais de 65% de probabilidade que todas as músicas listadas no Top 10 da Billboard daquele ano seriam um sucesso. Ainda hoje há especulações que artistas e estúdios contratam a Polyphonic HMI para melhorarem suas músicas.

2.3 *Word embeddings*

Word embeddings é uma técnica de processamento de linguagem natural na qual palavras ou frases são mapeadas em vetores numéricos densos, sendo capazes de capturar similaridades contextuais e semânticas entre elas. Diferentemente de outros métodos de codificação, que criam representações utilizando vetores esparsos com um alto número de dimensões, *word embeddings* criam vetores densos com um número reduzido de dimensões, geralmente variando entre 50 e 768. Este último é o caso do vetor usado pelos modelos BERT.

Uma das técnicas mais utilizadas de *word embeddings* é o Word2Vec, criado em 2013 por uma equipe liderada por Tomas Mikolov, no Google (Mikolov et al. 2013). A técnica era inovadora por utilizar uma rede neural de duas camadas para gerar vetores de palavras em um espaço de dimensão relativamente baixo, mantendo a riqueza semântica e sintática. São utilizados dois modelos principais: o *Continuous Bag of Words* (CBOW) e o *Skip-Grams*, que funcionam de forma complementar, sendo o primeiro capaz de prever uma palavra dado o contexto das palavras ao redor, enquanto o segundo é capaz de prever o contexto a partir da palavra, permitindo que se capture diversas relações linguísticas.

Outra técnica muito utilizada para representação de palavras em vetores numéricos é o TF-IDF (*term frequency - inverse document frequency*), uma abordagem estatística criada em 1972 por Karen Spärck Jones que calcula a importância de cada termo para cada documento em relação aos demais documentos. O primeiro termo, TF, calcula a importância da palavra para o documento em que está contida, enquanto o segundo termo, DF, calcula a importância da palavra entre todos os documentos. Multiplicando a frequência do termo (TF) pela frequência inversa do documento (IDF), obtemos a importância relativa daquela palavra. Palavras frequentes em muitos documentos terão sua importância reduzida, enquanto aquelas frequentes em poucos documentos terão sua importância aumentada, permitindo a extração das palavras mais importantes de cada

documento (Robertson 2004), tornando-se uma ferramenta poderosa na classificação de documentos, modelagem de tópicos e criação de modelos de busca.

$$TF\text{-}IDF(t, d, D) = TF(t, d) \times \log\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right) \tag{2.1}$$

onde TF(t, d) é a frequência do termo t no documento d,
|D| é o número total de documentos no corpus D, e
|{d ∈D : t ∈d}| é o número de documentos que contêm o termo t.

2.4 Redução de dimensionalidade com PCA

PCA (*Principal Component Analysis*), ou Análise de Componentes Principais, é uma técnica de aprendizado não supervisionado amplamente utilizada para redução de dimensionalidade, preservando o máximo possível da variabilidade dos dados. A técnica envolve a normalização dos dados, o cálculo da matriz de covariância, dos autovetores e autovalores, e a partir disso, calculam-se os componentes, que são combinações lineares das variáveis originais, com o objetivo de simplificar dados com muitas dimensões (Ma 2014). Neste estudo, a técnica de PCA foi aplicada para reduzir a dimensionalidade das representações vetoriais obtidas pelos algoritmos TF-IDF e Word2Vec baseadas nas transcrições de canções do gênero *rap*.

2.5 Análise textual de canções

A análise textual baseia-se em características extraídas das transcrições das canções, complementando a análise sonora. A partir das transcrições, é possível capturar os tópicos abordados, as classes gramaticais mais utilizadas, o teor e a formalidade do conteúdo transmitido e como essas variáveis interagem com a variável alvo. O estudo de (Fell e Sporleder 2014) conseguiu identificar as características mais relevantes para os usuários na diferenciação entre as canções consideradas melhores e piores por gênero musical, classificadas a partir da avaliação de usuários no site rateyourmusic.com[1]. Por exemplo, o uso de gírias é mais importante no gênero *rap* do que no gênero *pop/rock*. Por outro lado, canções do gênero *metal* com maior diversidade lexical tendem a ser mais apreciadas. A análise textual permite modelar as letras das canções de forma a atender cada vez mais aos desejos do público-alvo.

[1] https://rateyourmusic.com/

2.6 Aprendizado de Máquina e classificação binária

O Aprendizado de Máquina é o nome dado a um grande conjunto de algoritmos que, atualmente, são os mais usados nos sistemas de Inteligência Artificial (IA). Tais algoritmos permitem às máquinas aprender a partir dos dados. Durante o processo de aprendizado, as máquinas identificam padrões e erros, o que permite melhorar seu desempenho na tomada de decisões sem a necessidade de intervenção humana direta.

A classificação binária é uma técnica usada para categorizar dados em duas classes distintas. Alguns exemplos comuns são a identificação de *spams*, fraudes em transações bancárias, diagnósticos de doenças, entre outros. Neste estudo, a classificação binária será utilizada para prever se uma música pertence à categoria "Mais ouvidas"ou "Menos ouvidas".

Para avaliar os modelos de classificação binária, podemos utilizar a acurácia (proporção de previsões corretas, tanto positivas quanto negativas, em relação ao total de previsões), precisão (proporção de previsões corretas de uma classe positiva em relação ao total de previsões positivas), cobertura (proporção de previsões corretas de uma classe positiva em relação a todos os exemplos relevantes dessa classe), Medida-F (ou F_1) (média harmônica entre precisão e cobertura, equilibrando ambas as métricas), entre outras métricas.

Para o cálculo das métricas mencionadas, podemos utilizar a matriz de confusão apresentada na imagem 2.1.

		Valor Predito	
		Mais ouvidas	**Menos ouvidas**
Valor Real	**Mais ouvidas**	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Menos ouvidas	Falso Positivo (FP)	Verdadeiro Negativo (FN)

Figura 2.1: Matriz de confusão utilizada para calcular as métricas de avaliação descritas, como acurácia, precisão, cobertura e medida-f.

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.2}$$

onde TP são verdadeiros positivos, TN são verdadeiros negativos,
FP são falsos positivos e FN são falsos negativos.

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{2.3}$$

onde TP são verdadeiros positivos e FP são falsos positivos.

$$\text{Cobertura} = \frac{TP}{TP + FN} \tag{2.4}$$

onde TP são verdadeiros positivos e FN são falsos negativos.

$$F_1 = 2 \times \frac{\text{Precisão} \times \text{Cobertura}}{\text{Precisão} + \text{Cobertura}} \tag{2.5}$$

onde F_1 é a média harmônica de Precisão e Cobertura.

2.7 Transformers

A arquitetura de redes neurais conhecida como *Transformers*, amplamente utilizada no campo do processamento de linguagem natural e introduzida pelo artigo (Vaswani et al. 2017), representa o estado da arte na criação de representações de palavras e frases em vetores numéricos, destacando-se pelo uso de mecanismos de atenção. Os mecanismos de atenção têm o papel de identificar de forma dinâmica as partes relevantes de uma entrada, adaptando-se ao contexto à medida que ele muda, diferentemente de técnicas como o Word2Vec, que cria uma representação estática das palavras. Assim, os *Transformers* geram várias representações para a mesma palavra em diferentes contextos, enquanto técnicas mais simples criam uma única representação por palavra.

Capítulo 3

Metodologia

Este capítulo tem como objetivo apresentar a metodologia aplicada nesta pesquisa. Inicialmente, é descrito o método utilizado para a coleta de dados e a definição da variável-alvo. Em seguida, detalhamos o processo de transformação dos dados, que inclui desde o processo de tokenização, remoção de *stop words*, transformação de contrações, cálculos de representações de palavras através de vetores numéricos, até o cálculo de métricas auxiliares que serão complementares às transcrições das canções. A seguinte seção aborda os conjuntos de dados resultantes das etapas anteriores, e por fim, é apresentado o processo de seleção de modelos a partir de um processo de busca exaustiva.

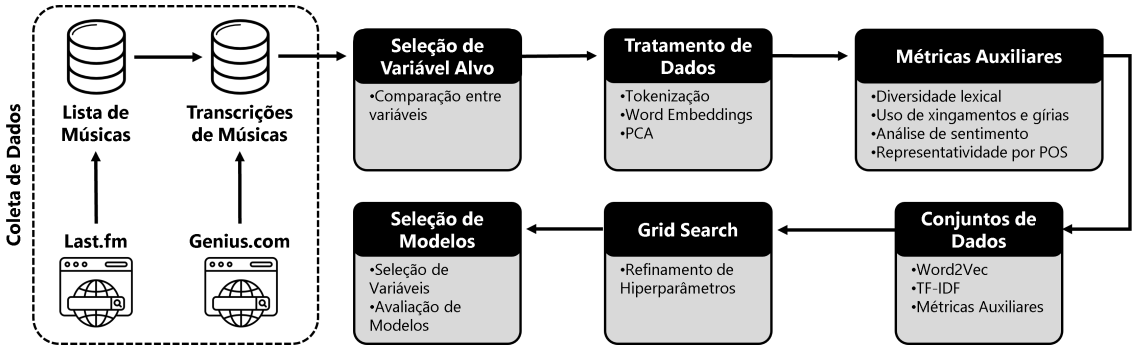

Figura 3.1: Diagrama dos processos descritos na metodologia.

3.1 Coleta de Dados

Para realização deste estudo, foi necessário obter um conjunto de dados previamente rotulados para a identificação de fatores relevantes na identificação de canções do gênero *rap* com potencial de sucesso comercial. Para isso, foi desenvolvida uma ferramenta de

raspagem de dados que fez a coleta de canções populares dentro do gênero. Os dados raspados têm origem em dois portais de internet chamados Last.fm[1] e Genius.com[2].

Primeiro foi utilizado o portal Last.fm, criado em 2002 como uma rádio web comunitária com o propósito de reproduzir músicas enquanto o usuário visualiza sua transcrição e interage com outros usuários, para identificar as canções mais populares dentro de um gênero específico.

Após aplicar o filtro ao gênero musical *rap*, foram coletadas "nome", "artista", "link para a transcrição da canção", "número de reproduções"e "número de ouvintes únicos"para as 10.000 canções mais populares dentro do gênero do *rap* e ordenadas por número de reproduções e ouvintes únicos.

Após o processo de raspagem, foram selecionadas as 1.000 primeiras canções mais ouvidas e categorizadas como tal, assim como foram selecionadas as 1.000 canções menos ouvidas, que foram categorizadas da mesma maneira, formando a lista de canções contidas no conjunto de dados analisado.

Para a segunda etapa de coleta de dados, foi utilizada a API do portal de internet Genius.com, criado em 2009 como um portal focado em canções do gênero *rap*, e que hoje é considerado o maior portal de música do mundo, para coletar as transcrições das 2.000 canções selecionadas na etapa anterior.

Além disso, 26 canções não foram encontradas através da API e foram desconsideradas.

Foi utilizado um modelo de linguagem pré-treinado da biblioteca spaCy[3] para identificar qual o idioma da música. Como a grande maioria das canções analisadas está no idioma inglês (Inglês: 1944, Alemão: 7, Espanhol: 5, Outros: 18), as canções nos demais idiomas foram desconsideradas para melhorar a precisão das análises. Por fim, foram removidas as canções duplicadas (por motivos de participação especial de outros artistas, remix ou redundância na plataforma) restando 1.915 transcrições. Dessas, 968 (50,5%) foram categorizadas como "Mais Ouvidas"e 947 (49,5%) como "Menos Ouvidas".

3.2 Seleção da variável-alvo

Nossa variável-alvo é calculada a partir da posição relativa da música em nossa lista ordenada de canções. Como mencionado na seção anterior, duas variáveis podem ser usadas para ordenar a lista de canções e assim definir sua classificação: o número de reproduções e número de ouvintes únicos. Ambas as variáveis foram testadas em todo o processo descrito na metodologia, e a variável que se revelou mais eficaz foi a de número de reproduções. As razões da escolha são as que seguem.

[1] https://www.last.fm/
[2] https://genius.com/
[3] https://spacy.io/universe/project/spacy_cld

A análise indica que algumas canções podem gerar interesse do público por motivos de marketing, polêmicas, participação em filmes e seriados, ou outros assuntos envolvendo o artista. Entretanto, esse interesse não se mantém, pois a música não possui a qualidade necessária para manter os ouvintes engajados, e, com isso, ela possui um alto número de ouvintes, mas um baixo número de reproduções. Por outro lado, ao utilizarmos o número de reproduções podemos diferenciar quais canções engajam mais o público de forma contínua, sendo o objetivo dessa pesquisa.

Os modelos que fizeram uso da variável-alvo calculada a partir da lista de canções ordenadas pelo número de reproduções tiveram uma melhora significativa em relação aos modelos que fizeram uso da variável-alvo calculada a partir do número de ouvintes únicos, tendo uma diferença de até 11 pontos percentuais em suas métricas de avaliação.

3.3 Tratamento de Dados

Abaixo temos o processo aplicado para a transformação das transcrições das letras das canções em listas de *tokens* e de representações por vetores numéricos, assim como o detalhamento das métricas auxiliares.

3.3.1 *Tokenização*

Para realizar a *tokenização* das transcrições, foi criada uma função que executa as seguintes etapas:

- Para cada transcrição, é criado um objeto da biblioteca *Spacy* onde é possível extrair o *token*, o lema, o tipo da palavra, sua classe gramatical assim como outras características.
- Foram removidas as palavras identificas pelas classes gramaticais *PUNCT* (pontuações), *SPACE* (espaçamento), *SYM* (símbolos), *NUM* (números) e *X* (outros).
- Também foram removidas as palavras contidas na lista padrão de *stop words* da biblioteca *Spacy*. Foram adicionadas as palavras *"feat"*, *"ft"* e *"embed"* por serem resíduos do processo de raspagem dos dados.
- Foram removidas palavras com apenas um caractere, a fim de reduzir erros de transcrição.

- É utilizada a biblioteca *contractions*[4] para converter as contrações da língua inglesa, que normalmente ocorrem ao se combinar um verbo com sua negação. (ex.: *have not = haven't, did not = didn't*)

- Para palavras que foram abreviadas ou escritas de uma forma diferente, seja por erros na transcrição ou por conta de uma linguagem mais informal, foram aplicadas correções para facilitar o entendimento e agrupamento das mesmas. Foi criado um dicionário com 1.146 palavras distintas para aplicar as correções de forma consistente (ex.: *runnin = running, eatin = eating*).

Por fim, são extraídos os lemas das palavras, isto é, sua base ou forma canônica, a fim de melhorar o agrupamento das palavras e facilitar o entendimento do contexto de cada música.

A fim de ilustrar o resultado do pré-processamento, segue um trecho da música *Empire State of Mind* do rapper JAYZ antes e depois das transformações.

Antes: *Welcome to the meltin' pot, corners where we sellin' rock*

Depois: *welcome melting pot corner sell rock*

3.3.2 Representação por vetores numéricos

Foram utilizadas duas técnicas de vetores de palavras para representar as transcrições das canções para os nossos modelos. São elas:

- Word2Vec: O algoritmo escolhido foi o CBOW, que possui uma janela pela qual se tenta prever a palavra atual a partir das palavras ao redor, tendo como diferencial a captura do contexto. A seguir temos os parâmetros utilizados:
  - Tamanho do vetor (vector_size) = 300
  - Janela de contexto (window) = 3
  - Contagem mínima de ocorrências (min_count) = 5

- TF-IDF:
  - Contagem mínima de ocorrência (min_df) = 5
  - Combinações de ngramas (ngram_range) = (1, 2)

[4] https://github.com/kootenpv/contractions

3.3.3 Redução de Dimensionalidade

Foi utilizado PCA como técnica de redução de dimensionalidade, aplicado aos dados do TF-IDF, a fim de reduzir a complexidade dos modelos e ruídos, mantendo as características mais significativas.

A transformação dos dados através do PCA começa pela normalização dos dados para que tenham média zero e desvio-padrão 1. Em seguida, calcula-se a matriz de covariância, e a partir dela, os autovetores e autovalores, que representam as direções de maior variação nos dados e a magnitude dessas variações respectivamente. Por fim, os dados são projetados nessas novas direções, o que permite reduzir a dimensionalidade enquanto se preserva a maior parte da variância original dos dados.

Foram testados dois conjuntos de dados, sendo o primeiro composto pelos primeiros 1.010 componentes, representando 80% da variância explicada dos dados, enquanto o segundo inclui apenas os primeiros 100 componentes.

3.3.4 Métricas Auxiliares

Além das transcrições, calcularam-se métricas auxiliares para testar formas alternativas de identificar as características que mais atraem ouvintes.

Uso de Xingamentos e Gírias

Para identificar se o uso de xingamentos ou gírias influencia no interesse do público, foram criados três dicionários distintos.

- Um dicionário estático contendo os xingamentos mais comuns da língua inglesa.
- Um dicionário construído a partir do portal de internet *Wiktionary*[5], onde é possível coletar o significado, classe gramatical e demais características para cada palavra.
- Um dicionário construído a partir do portal de internet *Urbandict*[6], especializado em gírias urbanas.

Para identificar se uma palavra é um xingamento, verifica-se se ela está contida no dicionário de xingamentos. Para identificar se uma palavra é uma gíria, verifica-se se ela está contida no dicionário de palavras extraídas do *Urbandict* e não no dicionário de palavras comuns. Não podemos simplesmente considerar as palavras contidas no *Urbandict* como sendo gírias, pois os usuários utilizam a plataforma de forma irregular e

[5] https://www.wiktionary.org/
[6] https://www.urbandictionary.com/

atribuem significados com teor humorístico a palavras comuns, restringindo a identificação de gírias somente para palavras específicas e que não estejam contidas em dicionários comuns. Essa abordagem foi sugerida pelo artigo (Fell e Sporleder 2014), entretanto, não leva em consideração o contexto das palavras, sendo uma limitação da técnica. As demais palavras são consideradas como palavras comuns.

Diversidade Lexical

Foi calculado o número de lemas (*tokens*) e frases contidas em cada música (*sentences*), assim como a contagem de lemas únicos (*types*). A partir do número de lemas únicos na música e do número total de lemas, é possível calcular o TTR (*type-token ratio*), que representa a diversidade lexical. Também foi calculado o número médio de lemas por frase (*lemmas_per_sentence*).

Análise de Sentimento

Através da biblioteca *SpacyTextBlob*[7], foram calculadas a polaridade, representando a orientação emocional de um texto, que varia de -1 (negativo) a 1 (positivo), e a subjetividade, referindo-se ao grau em que um texto contém opiniões, emoções ou julgamentos pessoais, variando de 0 (pouca subjetividade) a 1 (muita subjetividade). Para ambas as variáveis, são utilizados modelos pré-treinados.

Representatividade por Classe Gramatical

Também foi calculada a representatividade de cada parte da oração (em inglês, *part of speech* (POS)) em cada música, a fim de identificar se a maior utilização de alguma classe tem impacto na escolha da música pelos ouvintes.

Padronização

Para padronização, foi utilizada a normalização pelo teste-z (escore padronizado), ou seja, a transformação dos dados para média 0 e desvio padrão 1, com o objetivo de padronizar escalas distintas e facilitar a comparação entre as canções. As variáveis transformadas foram:

- Número de lemas
- Número de lemas únicos
- Número de frases

[7] https://spacy.io/universe/project/spacy-textblob

- Número de lemas por frase
- Número de xingamentos
- Número de gírias

O cálculo matemático utilizado para a transformação pode ser observado abaixo.

$$z = \frac{(x - \mu)}{\sigma}$$

onde:

z : Valor padronizado resultante após a normalização.
x : Valor original da variável antes da normalização.
μ : Média dos valores originais da variável.
σ : Desvio padrão dos valores originais da variável.

3.4 Conjuntos de Dados

Para os experimentos, foram utilizados os seguintes conjuntos de dados:

- Word2Vec: Representações vetoriais das transcrições das canções se utilizando do algoritmo CBOW. Como cada palavra possui um vetor próprio, calculou-se um vetor médio para representar cada música.
- TF-IDF: Representações vetoriais das transcrições das canções se utilizando do algoritmo TF-IDF.
- PCA(TF-IDF)(80pct): Técnica de redução de dimensionalidade PCA aplicada ao conjunto de dados TF-IDF. Foram selecionados os N primeiros componentes que explicam 80% da variância dos dados.
- Métricas auxiliares: Características mencionadas na subseção "Métricas Auxiliares".
- Métricas auxiliares (FS): Características mencionadas na subseção "Métricas Auxiliares", após um processo iterativo de seleção de variáveis, utilizando-se do algoritmo de regressão logística. Para cada iteração, remove-se a variável com o p-valor maior que 0,05. O processo termina quando todas as variáveis restantes possuem p-valor menor ou igual a 0,05.

- Métricas auxiliares + PCA(TF-IDF)(n=100): Combinação das características mencionadas na subseção "Métricas Auxiliares"com os 100 primeiros componentes do PCA aplicado ao objeto TF-IDF.

3.5 Seleção de Modelos

A pesquisa foi realizada para 5 diferentes tipos de modelos, que são: kvizinhos mais próximos (KNN), Regressão logística, Máquina de vetores de suporte (SVM), Árvore de decisão e Floresta Aleatória.

Em todos os experimentos, os dados foram divididos em 75% para treinamento e 25% para teste.

Também foi utilizada uma técnica de *GridSearch* para encontrar os melhores hiperparâmetros para cada modelo, sendo o universo testado representado a seguir:

- k-vizinhos mais próximos (KNN)
  - Número de vizinhos (n_neighbors): 3, 5, 7, 9
- Regressão logística
  - Penalidade (penalty): l1, l2
  - Inverso da força da regularização (C): 0.01, 0.1, 1, 10, 100
  - Intercepto (fit_intercept): True, False
  - Pesos das classes (class_weight): None, balanced
- Máquina de vetores de suporte
  - Inverso da força da regularização (C): 0.1, 1, 10
  - Kernel (kernel): linear, rbf, poly, sigmoid
  - Coeficiente do Kernel (gamma): scale, auto
  - Grau (degree): 2, 3, 4
- Árvore de decisão
  - Profundidade máxima (max_depth): 3, 5, 10, 20, 30
  - Mínimo de amostras para divisão (min_samples_split): 2, 5, 10
  - mínimo de amostras por folha (min_samples_leaf): 1, 2, 4
- Floresta Aleatória

- Número de estimadores (n_estimators): 100, 200, 300
- Profundidade máxima (max_depth): 5, 8, 10
- Mínimo de amostras para divisão (min_samples_split): 2, 5, 10
- mínimo de amostras por folha (min_samples_leaf): 1, 2, 4
- Utilização de bootstrap (bootstrap): True, False

Capítulo 4

Resultados e Discussão

Neste capítulo são abordados os resultados obtidos por combinações de modelos, técnicas e conjuntos de dados mencionados anteriormente. Os resultados são explorados com o objetivo de encontrarmos formas consistentes de identificar canções do gênero *rap* com potencial comercial.

4.1 Análise por representações vetoriais

As duas técnicas de representações vetoriais empregadas, apesar de distintas, têm como o mesmo objetivo capturar o significado e importância dos termos contidos nas transcrições das canções. Nesta seção são apresentados os resultados obtidos para cada conjunto de dados, bem como se discute qual técnica é mais adequada para esta pesquisa.

4.1.1 Visualização dos dados

Após o processo de limpeza e pré-processamento dos dados, identificou-se a distribuição de frequência de palavras, conforme ilustrado na Figura 4.1.

Cerca de 47,9% das palavras ocorrem somente uma vez, fenômeno também conhecido como *Hápax Legômena*. Após uma análise exaustiva, identificou-se que, ao limitar o conjunto de dados a palavras com frequência maior ou igual a 5, obtêm-se melhores resultados na identificação das canções com maior potencial de sucesso. O provável motivo é que essa abordagem remove as palavras mais comuns e realça as diferenças entre as canções, diminuindo o ruído e reduzindo a complexidade dos modelos. Por esse motivo, ambas as técnicas de representação por vetores numéricos (Word2Vec e TF-IDF) consideram apenas termos com frequência maior ou igual a 5.

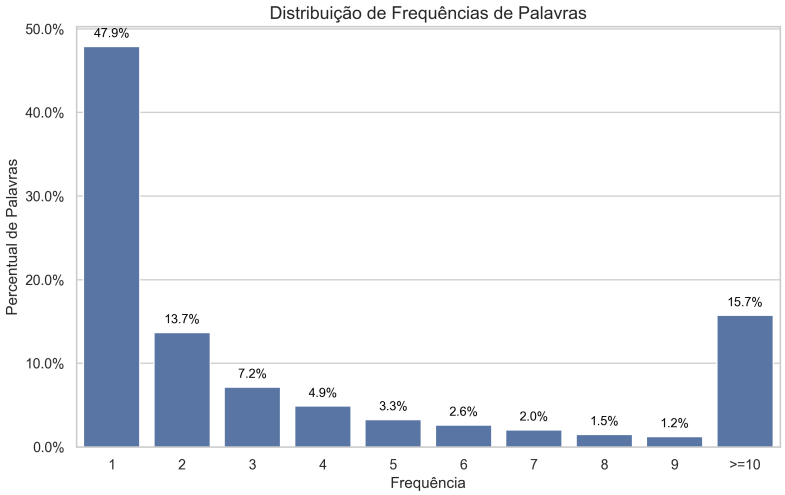

Figura 4.1: Distribuição de frequência de palavras após a etapa de limpeza e pré-processamento dos dados.

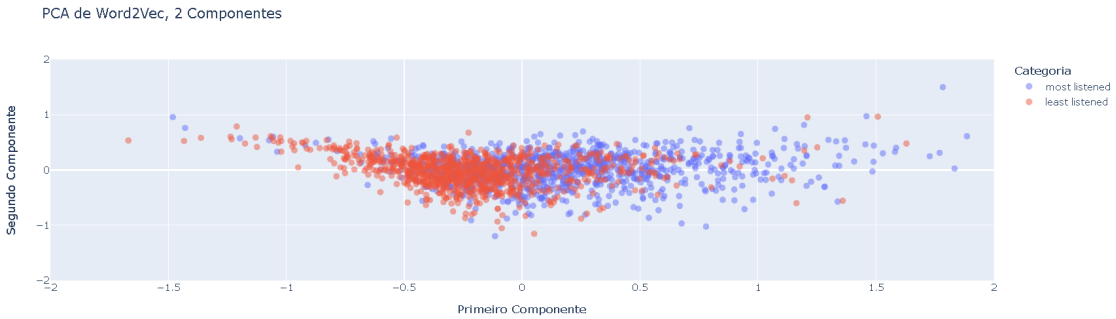

Figura 4.2: Representação 2D dos componentes extraídos a partir da representação vetorial gerada pelo algorítimo Word2Vec.

Ao aplicar PCA às representações vetoriais, é possível ter uma ideia de como os dados estão distribuídos entre as duas categorias (Mais Ouvidas e Menos Ouvidas), observando a distribuição dos dados em 2 e 3 componentes.

Nas Figuras 4.2 e 4.3, observam-se as representações 2D e 3D dos componentes extraídos a partir da representação vetorial gerada pelo algoritmo Word2Vec. Se as representações apresentassem muita sobreposição entre as duas categorias analisadas ou um padrão muito complexo, isso indicaria a necessidade de utilizar algoritmos mais complexos para diferenciar as duas classes.

Em seguida, nas figuras 4.4 e 4.5 observam-se as representações 2D e 3D dos componentes extraídos a partir da representação vetorial gerada pelo algoritmo TF-IDF. Novamente, nota-se uma boa separação entre as duas categorias analisadas.

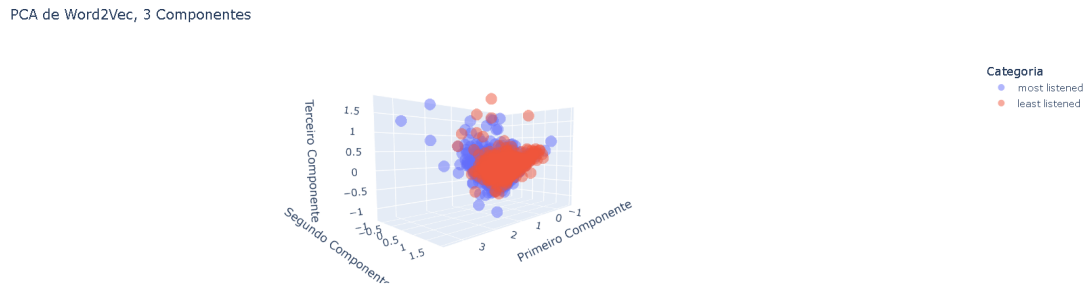

Figura 4.3: Representação 3D dos componentes extraídos a partir da representação vetorial gerada pelo algorítimo Word2Vec.

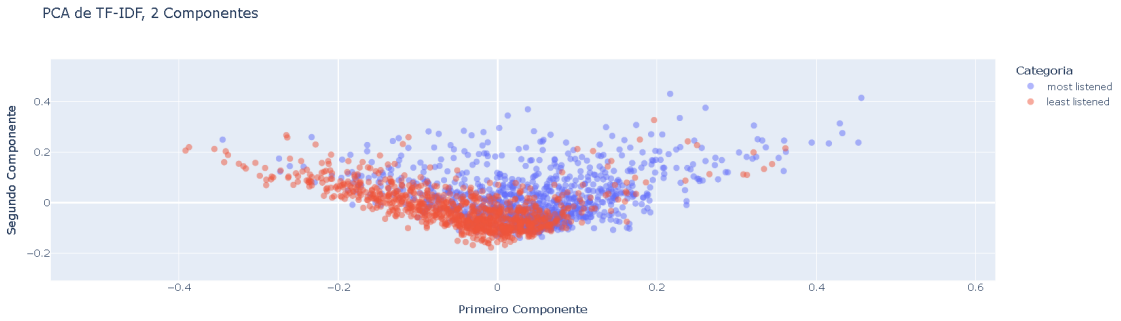

Figura 4.4: Representação 2D dos componentes extraídos a partir da representação vetorial gerada pelo algorítimo TF-IDF.

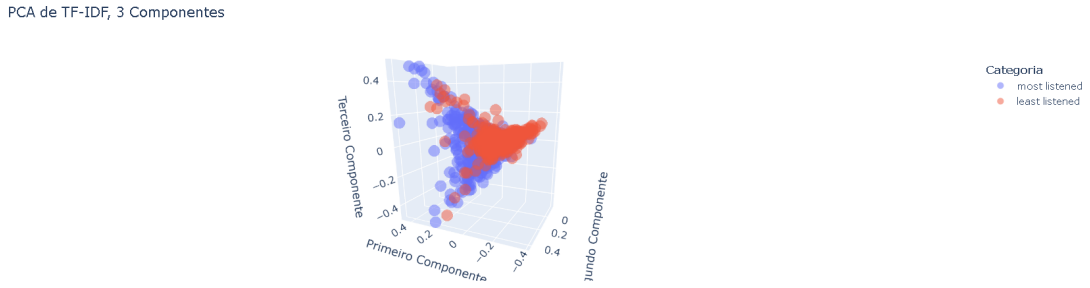

Figura 4.5: Representação 3D dos componentes extraídos a partir da representação vetorial gerada pelo algorítimo TF-IDF.

As imagens demonstram que é possível realizar uma boa separação entre as duas categorias usando apenas duas ou três dimensões, sugerindo que, com o uso de mais dimensões, os modelos poderão separar as categorias ainda mais facilmente.

4.1.2 Categorização

Word2Vec

Na tabela 4.1 os modelos são apresentados em ordem de acurácia, após um processo iterativo de *grid search* para encontrar os melhores parâmetros de cada um.

Modelo	Acurácia	Medida-F	Cobertura	Precisão
Floresta Aleatória	**72,44**	**72,27**	**72,44**	72.47
KNN	70.35	70.19	70.35	70.33
SVM	70.35	68.32	70.35	**74,26**
Regressão Logística	59.08	50.86	59.08	66.88
Árvore de Decisão	56.79	56.31	56.79	56.45

Tabela 4.1: Métricas de Desempenho dos modelos utilizando Word2Vec.

Observa-se que o melhor modelo foi o Floresta Aleatória, enquanto o de Regressão Logística ficou em segundo pior, sugerindo que os dados são mais complexos e não lineares. A Floresta Aleatória tem vantagem sobre as Árvores de Decisão justamente por lidar melhor com o sobreajuste, já que cria várias árvores com diferentes parâmetros e conjuntos de dados, sendo um método de *ensemble* (combinação de vários modelos).

TF-IDF

Na tabela 4.2 os modelos são apresentados em ordem de acurácia, após um processo iterativo de *grid search* para encontrar os melhores parâmetros de cada um.

Modelo	Acurácia	Medida-F	Cobertura	Precisão
Floresta Aleatória	**81,84**	**81,84**	**81,84**	**81,86**
Regressão Logística	81.42	81.44	81.42	81.49
SVM	80.58	80.60	80.58	80.68
KNN	76.41	76.44	76.41	76.62
Árvore de Decisão	70.56	70.41	70.56	70.53

Tabela 4.2: Métricas de Desempenho dos modelos utilizando TF-IDF.

Novamente, observa-se que o modelo de Floresta Aleatória alcançou a maior acurácia, enquanto o segundo melhor, a Regressão Logística, apresentou uma diferença inferior a 1 ponto percentual em relação ao primeiro, contrastando com o experimento anterior.

Observa-se, também, que a diferença entre todos os modelos diminuiu consideravelmente, e o primeiro colocado obteve uma acurácia aproximadamente 9 pontos percentuais maior em comparação ao melhor modelo treinado com a representação gerada pelo Word2Vec. Os dados sugerem que o TF-IDF conseguiu gerar uma representação vetorial superior à do Word2Vec, além de possibilitar uma representação mais linear, o que é evidenciado pela melhora significativa no modelo de Regressão Logística. O principal motivo para a representação gerada pelo Word2Vec não ter sido tão eficiente pode ser atribuído ao tamanho do corpus utilizado. Para lidar com essa limitação, é válida a experimentação com um modelo pré-treinado em um corpus maior, a fim de testar se haverá aumento relevante na acurácia dos modelos.

A tabela 4.3 nos mostra que, ao reduzir a dimensionalidade de nossos dados, é possível alcançar resultados bastante semelhantes aos obtidos com o conjunto de dados completo para alguns modelos. Isso é feito em troca de uma menor complexidade do modelo e de custos computacionais reduzidos. Para esse experimento foram utilizados os primeiros n componentes gerados a partir do TF-IDF, que explicam 80% da variância dos dados, resultando em uma redução inferior a um ponto percentual para o modelo de Regressão Logística.

Modelo	Acurácia	Medida-F	Cobertura	Precisão
Regressão Logística	**80,58**	**80,57**	**80,58**	**80,57**
SVM	80.17	80.17	80.17	80.17
KNN	78.50	78.52	78.50	78.85
Floresta Aleatória	74.74	74.73	74.74	74.72
Árvore de Decisão	72.23	72.25	72.23	72.26

Tabela 4.3: Métricas de Desempenho dos modelos utilizando os N primeiros componentes gerados a partir do TF-IDF que explicam 80% da variância dos dados).

Métricas Auxiliares

O objetivo de utilizar as variáveis descritas na subseção "Métricas Auxiliares"foi não depender exclusivamente do conteúdo das transcrições das canções, mas também identificar outras características com capacidade preditiva na identificação de canções com potencial de sucesso no gênero *rap*. Na tabela 4.4 verifica-se que as métricas auxiliares possuem, de fato, capacidade preditiva.

O modelo Floresta Aleatória novamente apresentou a melhor performance. No entanto, devido à pequena diferença em relação ao segundo colocado e a fim de melhorar a interpretabilidade das variáveis utilizadas, optou-se por selecionar o modelo de Regressão Logística como o principal.

Modelo	Acurácia	Medida-F	Cobertura	Precisão
Floresta Aleatória	**76,20**	**76,21**	**76,20**	**76,24**
Regressão Logística	75.37	75.40	75.37	75.74
SVM	75.37	75.38	75.37	75.99
Árvore de Decisão	68.06	68.07	68.06	68.73
KNN	63.47	63.49	63.47	63.53

Tabela 4.4: Métricas de Desempenho dos modelos utilizando as métricas auxiliares.

Após um processo iterativo de seleção de variáveis, em que, a cada iteração, foi removida uma variável com p-valor maior que 0,05, até que todas as variáveis restantes tivessem p-valor menor ou igual a 0,05.

Modelo	Acurácia	Medida-F	Cobertura	Precisão
Regressão Logística (FS)	**77,45**	**77,48**	**77,45**	**77,57**
Floresta Aleatória	76.20	76.21	76.20	76.24
Regressão Logística	75.37	75.40	75.37	75.74

Tabela 4.5: Comparativo do modelo de Regressão Logística antes e após o processo de seleção de variáveis.

Observa-se na tabela 4.5 que o processo de seleção de variáveis foi benéfico, melhorando a acurácia do modelo em aproximadamente 2 pontos percentuais, superando assim o modelo de Floresta Aleatória, que anteriormente ocupava a primeira posição.

Ao todo, obtivemos dez variáveis resultantes do processo iterativo de seleção, como observado na tabela 4.6.

Variável	Coeficiente	DP	z	$p > \|z\|$
INTJ	15.7638	3.4775	4.5331	>0.0001
ADV	6.6539	3.1111	2.1388	0.0325
PRON	4.7202	1.3827	3.4138	0.0006
Subjectivity	1.6984	0.6277	2.7059	0.0068
Slang	0.2302	0.0924	2.4920	0.0127
Types	−0.4079	0.1305	−3.1257	0.0018
Words per Sentence	−0.7872	0.1927	−4.0841	>0.0001
PROPN	−3.1665	1.4990	−2.1123	0.0347
NOUN	−6.6993	1.1392	−5.8809	>0.0001
CCONJ	−38.5557	5.1183	−7.5329	>0.0001

Tabela 4.6: Lista de variáveis restantes após o processo de seleção de variáveis e seus coeficientes.

As três variáveis com os maiores coeficientes positivos, ou seja, aquelas cujos valores mais elevados *aumentam* a probabilidade de uma música ser classificada como "Mais ouvidas", foram:

1. Interjeições (INTJ), palavras ou expressões que expressam emoção ou exclamação.
   i.e. *"Wow!", "Ouch!", "Hey!"*

2. Advérbios (ADV), modifica verbos, adjetivos ou outros advérbios, geralmente indicando maneira, tempo, lugar, frequência ou grau.
   i.e. *"quickly", "yesterday", "very"*

3. Pronome (PRON), substitui substantivos ou outras palavras nominais para evitar repetições, produzir generalizações ou manter anonimato.
   i.e. *"he", "it", "who"*

Por outro lado, as variáveis que apresentam um efeito inversamente proporcional, isto é, aquelas cujos valores mais elevados *diminuem* a probabilidade de uma música ser classificada como "Mais ouvidas", foram:

1. Conjunções coordenativas (CCONJ), liga palavras, frases ou cláusulas de igual importância gramatical na frase.
   i.e. *"and", "but", "or"*

2. Substantivos (NOUN), nomeia pessoas, lugares, coisas, ideias ou conceitos. Podem atuar como sujeito ou objeto em uma frase.
   i.e. *"dog", "city", "happiness"*

3. Substantivo Próprio (PROPN), nomes específicos de pessoas, lugares e organizações.
   i.e. *"Brazil", "New York", "Amazon"*

canções que incorporam interjeições, advérbios e pronomes tendem a ser mais populares, sugerindo uma preferência do público por canções emocionalmente expressivas e pessoais. Por outro lado, o uso frequente de conjunções coordenativas, substantivos comuns e próprios tem um efeito contrário, indicando que canções com orações mais longas, caracterizadas pelo uso de conjunções coordenativas, atraem menos o público. Essa tendência sugere que a simplicidade e a expressão direta de emoções são as melhores formas de engajar um público amplo.

Combinações desconsideradas

Os modelos que combinavam métricas auxiliares com representações vetoriais, tanto completas quanto parciais, foram desconsiderados devido à performance similar àquela dos modelos treinados apenas com representações vetoriais, não justificando, assim, um aumento em sua complexidade.

Ensemble

O último experimento combina os três melhores modelos em suas respectivas categorias de conjunto de dados, sendo eles:

- Word2Vec → Floresta Aleatória
- TF-IDF → Floresta Aleatória
- Métricas Auxiliares (FS) → Regressão Logística

A tabela 4.7 ilustra as métricas de desempenho obtidas pelo modelo combinado.

Modelo	Acurácia	Medida-F	Cobertura	Precisão
Ensemble	79.54	79.53	79.54	79.52

Tabela 4.7: Métricas de Desempenho do Modelo Ensemble utilizando características combinadas de TF-IDF, Word2Vec e Métricas Auxiliares.

Contrariamente às expectativas, os modelos apresentaram desempenho inferior quando combinados, sugerindo que as representações numéricas geradas pelo TF-IDF conseguem capturar as particularidades das canções de forma mais consistente quando utilizadas sozinhas, enquanto a utilização das métricas auxiliares é eficaz na identificação de características linguísticas relevantes para distinguir canções populares, orientando artistas ou produtores sobre quais direções tomar.

Capítulo 5

Conclusão

Esta pesquisa gerou modelos capazes de identificar canções do gênero *rap* com potencial comercial alcançando uma acurácia superior a 81%, usando como dados a representação numérica de suas transcrições combinadas com certas variáveis também associadas ao conteúdo linguístico, como o uso de classes gramaticais, xingamentos e gírias, diversidade lexical e análise de sentimento.

Identificou-se que canções mais simples e que empregam emoção direta e teor pessoal, isto é, incorporam interjeições, advérbios e pronomes, são mais eficazes em engajar o público.

Com um corpus de apenas 2.000 canções, foi possível identificar canções com potencial comercial com uma acurácia superior a 80%. Em comparação, uma pesquisa anterior realizada por (Fell e Sporleder 2014) utilizada como referência, empregou um corpus de aproximadamente 400 mil canções e alcançou cerca de 86% de acurácia na identificação de canções "boas"ou "ruins"com base na avaliação da audiência. Assim, com um corpus representando 0,5% dos dados empregados pela pesquisa de (Fell e Sporleder 2014), tivemos uma redução de apenas 6 pontos percentuais em acurácia. É importante destacar, entretanto, que as duas pesquisas, apesar de similares, têm objetivos distintos: a primeira foca no potencial comercial, medido pelo número de reproduções nas plataformas, enquanto a segunda leva em consideração a avaliação do público.

5.1 Limitações e Trabalhos Futuros

As principais limitações deste trabalho estão relacionadas a variáveis não capturadas, como ações de marketing e publicidade realizadas pelas produtoras, participação de músicas e artistas em séries, filmes e televisão, engajamento nas redes sociais e polêmicas associadas às músicas, entre outros fatores.

Uma possível melhoria seria trazida pela coleta de dados de redes sociais e plataformas de *streaming* para enriquecer a base de dados.

Outra limitação está associada ao pequeno tamanho do corpus utilizado, o que restringiu as análises ao uso de modelos de aprendizado de máquina menos complexos. Com um corpus maior, seria possível treinar modelos mais complexos ou realizar *fine tuning* em um modelo pré-treinado como os *transformers*, que representam o estado da arte no processamento de linguagem natural.

Referências

Duhigg, Charles (2012). *O Poder do Hábito*. Editora Objetiva.

Fell, Michael e Caroline Sporleder (ago. de 2014). "Lyrics-based Analysis and Classification of Music". Em: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Ed. por Junichi Tsujii e Jan Hajic. Dublin, Ireland: Dublin City University e Association for Computational Linguistics, pp. 620–631. URL: https://aclanthology.org/C14-1059.

Herremans, Dorien (2019). "Dance Hit Song Prediction". Em: *arXiv preprint arXiv:1905.08076*. URL: https://arxiv.org/abs/1905.08076.

Lourenço, Mariane Lemos (2010). "Arte, cultura e política: o Movimento Hip Hop e a constituição dos narradores urbanos". Em: *Psicol. Am. Lat.* 19. Acesso em 16 de janeiro de 2024. URL: http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1870-350X2010000100014&lng=pt&nrm=iso.

Ma, Yuan Zhe (fev. de 2014). "A Tutorial on Principal Component Analysis". Em: DOI: 10.13140/2.1.1593.1684.

Mikolov, Tomas et al. (2013). "Efficient Estimation of Word Representations in Vector Space". Em: *arXiv preprint arXiv:1301.3781*.

Miranda, J. H. D. A. (2016). "Relação de mercado e trabalho social no hip-hop". Em: *Cadernos Do CEAS: Revista crítica De Humanidades* 223, pp. 32–41. DOI: 10.25247/2447-861X.2006.n223.p32-41.

Neal, Meghan (2015). *A Machine Successfully Predicted the Hit Dance Songs of 2015*. https://www.vice.com/en/article/bmvxvm/a-machine-successfully-predicted-the-hit-dance-songs-of-2015. Accessed: 2024-01-05.

Robertson, Stephen (out. de 2004). "Understanding Inverse Document Frequency: On Theoretical Arguments for IDF". Em: *Journal of Documentation - J DOC* 60, pp. 503–520. DOI: 10.1108/00220410410560582.

Vaswani, Ashish et al. (2017). "Attention is all you need". Em: *arXiv preprint arXiv:1706.03762*.