LycirGen: Deep lerarning para geração de musicas

Felipe de Oliveira Santos
Pontifícia Universidade Católica de Campinas
felipe.os5@puccampinas.edu.br

Gabriel de Antonio Mazetto
Pontifícia Universidade Católica de Campinas
gabriel.am5@puccampinas.edu.br

Mateus Pereira Alves
Pontifícia Universidade Católica de Campinas
mateus.pa2@puccampinas.edu.br

Resumo

Este trabalho descreve o desenvolvimento e a avaliação de modelos de deep learning para a geração automática de letras de música, um processo que tradicionalmente exige alta criatividade e pode ser desafiador para artistas. Focando na reprodução e expansão de técnicas existentes, exploramos Redes Neurais Recorrentes Long Short-Term Memory (LSTM) Bidirecionais em nível de caractere. O objetivo principal foi implementar e avaliar modelos com diferentes estratégias de dropout e investigar o impacto do fine-tuning para capturar o estilo de um artista específico, utilizando Ariana Grande como estudo de caso. O dataset "Song Lyrics" do Kaggle foi pré-processado e dividido em conjuntos de treino, validação e teste, tanto para o treinamento inicial específico do artista quanto para um pré-treinamento em um corpus geral seguido de fine-tuning. As métricas de avaliação incluíram perdas, acurácia em nível de caractere, perplexidade, além de métricas textuais como Type-Token Ratio (TTR), Taxas de Repetição de N-gramas e Taxa de Existência de palavras. Os resultados demonstram que, embora os modelos iniciais já apresentassem capacidade de formar palavras e frases, o modelo fine-tuned (sem dropout) superou significativamente os demais, alcançando uma acurácia de teste de 0.7932 e gerando letras com maior coerência local e maior similaridade estilística ao artista alvo. Limitações como a coerência semântica global e o contexto de entrada restrito foram identificadas, sugerindo direções para trabalhos futuros, como a exploração de arquiteturas como CharacterBERT.

CCS Concepts

• Computing methodologies → Artificial intelligence → Natural language processing → Language models;

• Computing methodologies → Artificial intelligence → Machine learning → Machine learning approaches → Neural networks;

• Applied computing → Arts and humanities → Sound and music computing.

Keywords

Geração de Letras de Música; Deep Learning; LSTM Bidirecional; Processamento de Linguagem Natural; Geração de Texto em Nível de Caractere; Fine-Tuning; Lyric Generation.

1. INTRODUCÃO

A composição de letras de música é um processo intrinsecamente artístico, demandando um alto grau de criatividade e expressividade. No entanto, muitos artistas, tanto amadores quanto profissionais, frequentemente se deparam com desafios como bloqueios criativos ou limitações técnicas na verbalização de suas ideias musicais [1]. Nesse contexto, o Processamento de Linguagem Natural (PLN) e, mais especificamente, os modelos de deep learning, emergem como ferramentas promissoras. Eles oferecem novas formas de co-criação, funcionando como assistentes de inspiração e prototipação artística, auxiliando na superação desses obstáculos e na exploração de novas avenidas criativas [2][3].

O potencial impacto social dessas tecnologias é considerável, podendo democratizar o acesso à criação musical e fornecer aos artistas ferramentas que expandam suas capacidades expressivas [4].

1.1 Motivação

A principal motivação deste estudo reside na exploração da capacidade da Inteligência Artificial (IA) em auxiliar no processo criativo da composição musical. A escrita de letras, um componente fundamental da música, pode ser um processo árduo. A possibilidade de utilizar modelos de linguagem para gerar sugestões, completar ideias ou até mesmo criar rascunhos completos de letras pode significar uma economia de tempo e um estímulo à criatividade para compositores [5][6].

1.2 Aplicações Práticas e Trabalhos Relacionados

O campo da geração de letras de música com IA tem visto avanços significativos, com diversas ferramentas já disponíveis e moldando o futuro da composição. Exemplos notáveis incluem:

- LyricStudio: Uma plataforma que sugere trechos de letras com base em fragmentos iniciais fornecidos pelo usuário e suas preferências artísticas [7].
- Suno AI: Uma ferramenta que vai além da geração de letras, criando também o acompanhamento sonoro, oferecendo uma experiência de composição musical mais completa [8].

Este trabalho se baseia fundamentalmente no estudo de Ilakiyaselvan et al. [9], intitulado "Lyrics Generation Using LSTM and RNN". Este artigo explora a capacidade de Redes Neurais Recorrentes (RNNs) e, especificamente, LSTMs, para gerar letras de música personalizadas por artista, destacando como LSTMs bidirecionais podem capturar padrões de estilo e vocabulário mesmo em textos com estrutura sintática menos formal, como é comum em letras de música.

1.3 Objetivos do Trabalho

O presente trabalho tem como objetivo geral reproduzir e avaliar modelos de geração de letras musicais em nível de caractere utilizando redes Long Short-Term Memory (LSTM) bidirecionais. A investigação se concentra na análise do impacto de diferentes

estratégias de regularização e de especialização estilística, buscando compreender como essas abordagens influenciam a qualidade, coerência e expressividade das letras geradas automaticamente.

Para atingir esse objetivo geral, estabelecem-se quatro objetivos específicos. O primeiro consiste na preparação e pré-processamento de um dataset de letras de música. Esse processo inclui a segmentação dos textos em sequências de 100 caracteres, que serão utilizadas como unidades de entrada para o treinamento e a avaliação dos modelos, respeitando a abordagem de modelagem em nível de caractere.

O segundo objetivo visa implementar modelos LSTM bidirecionais com variações na aplicação da técnica de dropout, uma estratégia de regularização que busca reduzir o overfitting. Serão testadas três configurações: sem aplicação de dropout (modelo de linha de base), aplicação de dropout entre as camadas LSTM e aplicação de dropout antes da camada de saída totalmente conectada.

O terceiro objetivo consiste em avaliar os modelos gerados sob uma perspectiva tanto quantitativa quanto qualitativa. A análise quantitativa envolve métricas objetivas, como a diversidade lexical, enquanto a análise qualitativa considera aspectos subjetivos relacionados à coerência, fluidez e qualidade estética das letras produzidas.

Por fim, o quarto objetivo é investigar o efeito de uma estratégia de especialização de estilo por meio de fine-tuning. Para isso, um modelo LSTM bidirecional será inicialmente treinado em um corpus amplo e diversificado de letras musicais. Em seguida, será realizado o ajuste fino (fine-tuning) utilizando exclusivamente letras da artista Ariana Grande, com o intuito de adaptar o modelo para gerar letras que reflitam com maior fidelidade seu estilo particular de escrita e composição.

2. REFERENCIAL TEÓRICO

A geração automática de texto, especialmente no domínio criativo da música, apoia-se em avanços significativos em modelos de *deep learning* capazes de aprender padrões complexos a partir de grandes volumes de dados textuais.

2.1 Redes Neurais Recorrentes (RNNS) e Long Short-Term Mermory (LSTM)

Redes Neurais Recorrentes (RNNs) são uma classe de redes neurais projetadas para processar dados sequenciais, como texto ou séries temporais. Sua arquitetura permite que informações de passos anteriores na sequência influenciem o processamento dos passos atuais, através de conexões recorrentes que formam ciclos na rede. No entanto, RNNs simples sofrem com o problema do desaparecimento do gradiente (*vanishing gradient*), que dificulta o aprendizado de dependências de longo prazo em sequências extensas [10].

As redes Long Short-Term Memory (LSTM), introduzidas por Hochreiter & Schmidhuber (1997), são um tipo especializado de RNN projetado para mitigar esse problema. As LSTMs possuem uma estrutura de célula mais complexa, incorporando "*gates*" – *input gate*, *forget gate* e *output gate* – que controlam o fluxo de informação, permitindo que a rede aprenda quais informações armazenar, esquecer ou liberar ao longo da sequência. Essa capacidade torna as LSTMs particularmente eficazes para tarefas de modelagem de linguagem e geração de texto, onde o contexto de longo alcance é crucial [10][11].

2.2 LSTMs Bidirecionais

Em uma LSTM unidirecional padrão, a informação flui apenas em uma direção (do início para o fim da sequência). No entanto, para muitas tarefas de PLN, o contexto de palavras futuras pode ser tão importante quanto o de palavras passadas. As LSTMs Bidirecionais (BiLSTMs) abordam isso processando a sequência em duas direções: uma camada LSTM processa a sequência da esquerda para a direita, e outra camada LSTM processa da direita para a esquerda. As saídas dessas duas camadas são então concatenadas (ou combinadas de outra forma) em cada passo de tempo, fornecendo ao modelo uma representação mais rica do contexto [12][13]. No contexto da geração de letras, as BiLSTMs podem capturar melhor os padrões de estilo, vocabulário e estrutura frasal, como destacado por Ilakiyaselvan N. et al. (2023).

2.3 Dropout como Técnica de Regularização

O dropout é uma técnica de regularização amplamente utilizada em redes neurais para combater o overfitting. Durante o treinamento, o dropout "desliga" aleatoriamente uma fração dos neurônios (e suas conexões) em cada iteração. Isso impede que os neurônios coadaptem excessivamente e força a rede a aprender representações mais robustas e generalizáveis. Neste trabalho, exploramos diferentes estratégias de posicionamento do dropout com uma taxa de 0.2:

- **Sem Dropout:** Como linha de base.
- **Dropout entre camadas LSTM:** Aplicado às conexões entre as camadas LSTM bidirecionais.
- **Dropout antes da camada de saída densa:** Aplicado à saída concatenada das LSTMs antes de alimentar a camada *fully connected* final [14].

2.4 Geração de Texto em Nível de Caractere

A geração de texto pode ser abordada em diferentes níveis de granularidade, como palavra ou caractere. A modelagem em nível de caractere, utilizada neste estudo, envolve prever o próximo caractere em uma sequência com base nos caracteres anteriores. Embora possa exigir sequências de entrada mais longas para capturar o mesmo contexto que um modelo em nível de palavra, ela oferece vantagens significativas:

- **Vocabulário Reduzido:** O vocabulário é limitado ao conjunto de caracteres únicos (letras, números, pontuação), que é muito menor do que um vocabulário de palavras.
- **Tratamento de Palavras Raras e OOV:** Lida naturalmente com palavras raras, neologismos ou erros de digitação, pois pode construí-los caractere por caractere.
- **Captura de Nuances Estilísticas:** Pode aprender padrões subjacentes à grafia, uso de pontuação e formação de palavras que contribuem para o estilo de um autor ou gênero [15].

2.5 Fine-Tuning para Adaptação de Estilo

O fine-tuning é uma técnica de transfer learning onde um modelo pré-treinado em um dataset grande e geral é subsequentemente treinado (ou "ajustado") em um dataset menor e mais específico para uma tarefa particular. A ideia é que o modelo pré-treinado já aprendeu representações úteis da linguagem, que podem ser adaptadas para a nova tarefa com menos dados e tempo de treinamento [16]. No contexto deste trabalho, aplicamos uma estratégia de duas fases:

1. **Pré-treinamento:** Um modelo LSTM bidirecional é treinado em um corpus geral contendo letras de diversos artistas.
2. **Fine-Tuning:** O modelo pré-treinado é então ajustado utilizando apenas as letras de um artista específico (Ariana Grande), com o objetivo de especializar o modelo para gerar letras que mimetizem o estilo desse artista.

2.6 Estudo Base: "Lyrics Generetion Using LSTM and RNN"(Ilakiyaselvan N. et al., 2023)

O trabalho de Ilakiyaselvan N. et al. (2023) serve como principal referência teórica e metodológica para este projeto. Os autores exploram a capacidade das LSTMs para gerar letras de música específicas por gênero e artista, utilizando uma rede LSTM multicamadas com neurônios bidirecionais. Eles também investigam o impacto de diferentes posições de dropout e relatam a obtenção de letras de qualidade razoável. A arquitetura de 4 camadas LSTM bidirecionais e as estratégias de dropout exploradas neste estudo foram diretamente inspiradas por suas descobertas, buscando reproduzir e avaliar abordagens semelhantes em um contexto ligeiramente diferente e com um foco adicional no processo de fine-tuning para um artista específico.

3. MATERIAIS E MÉTODOS

Esta seção detalha o *dataset* utilizado, o pipeline de pré-processamento, a arquitetura do modelo, as estratégias de treinamento e as métricas empregadas para avaliação.

3.1 Dataset

O desenvolvimento e a avaliação dos modelos de geração de letras de música foram baseados no "Song Lyrics Dataset", uma coleção pública de letras disponível na plataforma Kaggle (Shah, 2021). Este dataset abrangente, contendo inicialmente 6027 músicas de variados artistas e gêneros, passou por um processo de limpeza que removeu letras vazias e *placeholders* (como "*lyrics for this song have yet to be released...*"), resultando em 5752 letras válidas para os propósitos do estudo. Para os experimentos focados em um artista específico, Ariana Grande foi selecionada como estudo de caso. De suas 308 letras originais, 14 foram descartadas por esses mesmos motivos, restando 294 letras limpas. Estas foram então divididas em conjuntos de treino (236 músicas), validação (29 músicas) e teste (29 músicas) para os modelos iniciais.

Com o objetivo de implementar uma estratégia de fine-tuning, foi utilizado o mesmo dataset previamente processado, porém com a exclusão de todas as músicas da artista Ariana Grande para a etapa de pré-treinamento do modelo geral. Esse corpus resultante, composto por 5458 letras, representou o conjunto geral de dados utilizados no pré-treinamento. A divisão desse dataset foi realizada da seguinte forma: 4366 músicas para treino, 545 para validação e 547 para teste. Após o pré-treinamento, as letras da Ariana Grande — previamente removidas — foram utilizadas exclusivamente na etapa de fine-tuning. Esse conjunto específico foi dividido em 236 músicas para treino, 29 para validação e 29 para teste, garantindo que o modelo fine-tuned fosse avaliado com dados nunca expostos durante o treinamento geral ou especializado.

3.2 Pipeline de Processamento e Pré-processamento de Dados

Um rigoroso pipeline de pré-processamento foi aplicado aos dados textuais para adequá-los à entrada dos modelos LSTM. O processo iniciou-se com a limpeza das letras, que envolveu a conversão integral do texto para letras minúsculas e a remoção de espaços em excesso, como tabulações e múltiplas quebras de linha. Seguiu-se uma filtragem de caracteres, onde foram preservados apenas caracteres alfanuméricos (a-z, 0-9) e um conjunto restrito de sinais de pontuação essenciais (.,!?;:'-); quaisquer outros símbolos ou caracteres especiais foram eliminados. As entradas que continham a frase *placeholder* indicativa de ausência de letra foram descartadas.

Após a etapa de limpeza, procedeu-se com a tokenização em nível de caractere. Um vocabulário foi construído a partir de todos os caracteres únicos presentes no *dataset* de treinamento relevante para cada fase (inicialmente do artista, depois do corpus geral). Este vocabulário consistiu consistentemente em 36 caracteres distintos. A cada caractere único no vocabulário foi atribuído um índice inteiro exclusivo.

Para treinar os modelos na tarefa de predição do próximo caractere, os textos das letras foram segmentados em múltiplas sequências de entrada (X) e seus correspondentes caracteres de saída (y). Adotou-se uma janela deslizante com um comprimento de sequência fixo em 100 caracteres. Desta forma, cada amostra de entrada era composta por 100 caracteres consecutivos, e o caractere alvo era aquele que se seguia imediatamente a essa sequência. A janela deslizava um caractere por vez ao longo de cada letra, maximizando a extração de amostras de treinamento. Este método gerou um volume expressivo de dados: o *dataset* inicial de Ariana Grande rendeu mais de 450 mil pares de dados, enquanto o corpus geral para pré-treinamento gerou mais de 9 milhões de sequências. Finalmente, os índices inteiros representando os caracteres foram utilizados como entrada para uma camada de *embedding* nos modelos, com uma dimensão de 256. Esta camada é responsável por aprender representações vetoriais densas e significativas para cada caractere, capturando relações contextuais e estilísticas.

3.3 Arquitetura do Modelo LSTM Bidirecional

A espinha dorsal de todos os modelos desenvolvidos neste trabalho foi uma rede LSTM (Long Short-Term Memory) bidirecional profunda, implementada utilizando a biblioteca PyTorch. O treinamento foi consideravelmente acelerado pelo uso de GPUs através da interface CUDA. A arquitetura detalhada do modelo compreende uma camada inicial de *Embedding*, que transforma os índices de caracteres de entrada em vetores densos com dimensão de 256. Sequencialmente, o modelo é composto por quatro camadas LSTM bidirecionais empilhadas. Cada uma

dessas camadas possui 256 unidades ocultas para o processamento na direção progressiva (*forward*) e outras 256 unidades para a direção regressiva (*backward*), resultando em um total de 512 unidades efetivas por camada LSTM bidirecional. A saída processada pela última camada LSTM, especificamente a representação do último passo de tempo da sequência de entrada, é então alimentada a uma camada densa (*fully connected*). Esta camada final projeta a representação aprendida para um vetor de dimensão igual ao tamanho do vocabulário de caracteres (36). Por fim, uma função de ativação Softmax é aplicada a este vetor de saída, gerando uma distribuição de probabilidade sobre todos os caracteres possíveis, indicando a predição do modelo para o próximo caractere da sequência.

3.4 Estratégias de Dropout

Com o intuito de investigar o impacto da regularização e mitigar o *overfitting*, três estratégias distintas de aplicação da técnica *dropout* foram exploradas, todas utilizando uma taxa de 0.2 quando ativadas. A primeira estratégia serviu como linha de base, consistindo em um modelo treinado completamente sem a aplicação de camadas de *dropout*. A segunda abordagem introduziu *dropout* após a saída de cada uma das camadas LSTM bidirecionais, antes que essa saída fosse passada como entrada para a camada LSTM subsequente. A terceira estratégia aplicou uma camada de *dropout* à saída concatenada proveniente da última camada LSTM bidirecional, especificamente no ponto que antecede a entrada para a camada final densa (*fully connected*) do modelo.

3.5 Processo de Treinamento

Nos modelos iniciais, que foram focados exclusivamente no *dataset* da artista Ariana Grande, o otimizador Adam foi empregado em conjunto com a função de perda CrossEntropyLoss, adequada para a tarefa de classificação multiclasse que é a predição do próximo caractere. O treinamento ocorreu ao longo de 10 épocas, com um tamanho de lote (*batch size*) de 128.

Para a abordagem de *fine-tuning*, uma estratégia de duas etapas foi adotada. Primeiramente, o modelo geral foi pré-treinado utilizando o "corpus geral", que compreende letras de uma variedade de artistas, excluindo-se, no entanto, os dados que seriam usados nas fases subsequentes de *fine-tuning* e teste da artista Ariana Grande. Este pré-treinamento estendeu-se por 3 épocas, com uma taxa de aprendizado de 0.001 e, crucialmente, foi realizado sem a aplicação de *dropout*, visando estabelecer uma base de conhecimento linguístico robusta. Na segunda etapa, o modelo pré-treinado teve seus pesos carregados e foi submetido ao processo de *fine-tuning*. Este ajuste fino foi realizado utilizando exclusivamente o conjunto de treino de Ariana Grande por 10 épocas. Para esta fase, uma taxa de aprendizado reduzida de 0.0001 foi utilizada, uma prática comum para permitir ajustes mais precisos nos pesos já pré-treinados. Similarmente ao pré-treinamento, o *fine-tuning* para os resultados finais reportados também foi conduzido sem *dropout*, com o mesmo tamanho de *batch size*. Em todas as fases de treinamento, o critério para salvar o melhor modelo foi a menor perda observada no respectivo conjunto de validação.

3.6 Métricas de Avaliação

A performance dos modelos de geração de letras foi aferida por meio de um conjunto diversificado de métricas quantitativas, abrangendo diferentes níveis de análise textual. No nível de caractere, as principais métricas incluíram a função de perda (*Cross-Entropy Loss*), que quantifica a divergência entre as distribuições de probabilidade preditas e as reais para o próximo caractere; e a acurácia na predição do próximo caractere, que indica o percentual de predições corretas;

Para avaliar a qualidade das letras geradas em sua totalidade, foi adotada uma abordagem em que, para cada música do conjunto de teste, os 100 primeiros caracteres da letra original foram utilizados como *seed*. A partir desse ponto, os modelos geraram texto caractere a caractere até atingir uma quantidade de palavras igual à da respectiva música de referência. Com isso, obteve-se um conjunto de letras geradas com tamanho semelhante ao do conjunto de teste real. Sobre essas letras, foram calculadas métricas textuais para análise quantitativa: o Type-Token Ratio (TTR), que mede a diversidade lexical; a Taxa de Repetição de Bigramas, que avalia a redundância de pares de palavras consecutivos; e a Taxa de Existência de Palavras, que calcula o percentual de palavras geradas que pertencem ao vocabulário original do corpus, indicando a aderência ao domínio e a geração de palavras plausíveis.

4. RESULTADOS E DISCUSSÃO

Nesta seção, são apresentados e discutidos os resultados obtidos nas diferentes fases de experimentação, abrangendo tanto as avaliações quantitativas quanto as qualitativas.

4.1 Motivação

Os primeiros experimentos focaram em treinar os modelos LSTM bidirecionais exclusivamente com o dataset de Ariana Grande, variando a estratégia de *dropout*.

Tabela 1: Métricas de Avaliação no Conjunto de Teste para Modelos Iniciais (Ariana Grande)

Métrica	**Sem Dropout**	**Dropout entre LSTMs**	**Dropout antes da FC**
Acurácia de Teste (Nível Caractere)	0.7288	0.6917	0.6858
Perda de Teste (Nível Caractere)	0.9172	1.0660	1.0780
TTR (Distinct-1 Ratio)	0.4961	0.4314	0.4449
Taxa de Repetição (Bigramas)	0.1568	0.2546	0.2448
Taxa de Existência de Palavras Únicas Geradas (%)	92.00	96.00	93.00

Fonte: Autoria Própria

Analisando os resultados, observa-se que o modelo treinado sem dropout apresentou o melhor desempenho em acurácia de caractere (0.7288) e as métricas mais favoráveis em termos de diversidade textual, com maior Type-Token Ratio (TTR = 0.4961) e menor taxa de repetição de bigramas (0.1568). Esses dados sugerem que, para este dataset específico e configuração

arquitetural, a aplicação de dropout não trouxe melhorias e pode, inclusive, ter prejudicado a modelagem de padrões estilísticos mais específicos da artista.

A Taxa de Existência de Palavras foi ligeiramente superior para o modelo com dropout entre camadas LSTM (96%), sugerindo uma maior aderência ao domínio linguístico aprendido, embora isso não tenha se traduzido em ganhos de coerência ou diversidade lexical.

Do ponto de vista qualitativo, as letras geradas por todos os modelos, ainda que contenham palavras corretas e algumas frases curtas bem formadas, frequentemente carecem de coerência sintática e semântica ao longo de trechos mais longos.

4.2 Impacto do Pré-treinamento Geral e Fine-Tuning

Concluídas as etapas de pré-treinamento do modelo geral e o fine-tuning com as letras da artista Ariana Grande, foi realizada uma comparação direta entre o melhor modelo inicial (sem *dropout*) e o modelo ajustado. A Tabela 2 apresenta os resultados dessa análise, destacando melhorias expressivas em desempenho e qualidade textual por parte do modelo *fine-tuned*.

Tabela 2: Comparativo de Métricas: Modelos Iniciais vs. Modelo Fine-tuned

Métrica	**Modelo Inicial (Sem Dropout)**	**Fine-tuned**
Acurácia de Teste (Nível Caractere)	0.7288	0.7932
Perda de Teste (Nível Caractere)	0.9172	0.7528
TTR (Distinct-1 Ratio)	0.4961	0.8900
Taxa de Repetição (Bigramas)	0.1568	0.0174
Taxa de Existência de Palavras Únicas Geradas (%)	92.00	98.00

Fonte: Autoria Própria

Os resultados indicam ganhos substanciais com o *fine-tuning*. A acurácia em nível de caractere subiu para 0.7932, demonstrando um modelo preciso na geração de texto.

As melhorias mais marcantes, no entanto, foram observadas nas métricas relacionadas à diversidade e à naturalidade textual. O Type-Token Ratio (TTR) quase dobrou, chegando a 0.8900, revelando um vocabulário mais variado. A taxa de repetição de bigramas despencou para apenas 0.0174, refletindo uma redução significativa na redundância. Além disso, a taxa de existência de palavras únicas geradas atingiu 98%, indicando maior aderência ao vocabulário original do corpus da artista.

Esses resultados evidenciam a eficácia do fine-tuning em especializar o modelo para capturar os traços estilísticos de um artista específico, mesmo utilizando uma arquitetura relativamente simples. A combinação de pré-treinamento amplo seguido de ajuste direcionado mostrou-se vantajosa tanto em termos de desempenho técnico quanto de riqueza linguística.

4.3 Análise Qualitativa da Geração de Letras

A comparação qualitativa das letras geradas é fundamental para ilustrar o impacto do processo de fine-tuning no desempenho do modelo. Para esta análise, foi utilizada como seed (texto inicial) a seguinte sequência: *"don't want nobody else around me just need you right here you're like the only thing that i see it's"*. Esse trecho corresponde ao início da música *"supernatural"* da artista Ariana Grande, lançada em 2024. Importante destacar que essa canção não estava presente na base de dados utilizada para o treinamento, o que torna o exemplo especialmente relevante para avaliar a capacidade de generalização estilística dos modelos.

O modelo inicial, treinado exclusivamente com os dados de Ariana Grande e sem a aplicação de dropout, gerou o seguinte trecho: *"to you ou does it just like it when we were the one you to know that we got that's a liw can my soul is christmas is yo"*. Em uma tradução aproximada, teríamos: *"para você ou será que é como se nós fôssemos o único que você soubesse que nós conseguimos isso é um liw pode minha alma é natal é você"*. Este exemplo, embora contenha algumas palavras reconhecíveis e estruturadas, apresenta uma organização frasal desconexa e carece de coerência semântica. Além disso, observa-se a geração de termos inexistentes no vocabulário da língua inglesa.

Em contraste, o modelo que passou pelo processo de pré-treinamento em um corpus geral e, subsequentemente, por fine-tuning com os dados de Ariana Grande (também sem dropout na fase de fine-tuning), produziu: *"love i just want to break your heart right back yeah all this time i was blind running 'round telling everybody my baby"*. (Tradução: *"amor eu só quero partir seu coração de volta yeah todo esse tempo eu estive cego correndo por aí dizendo a todo mundo meu bebê"*). Este trecho, apesar de não ser perfeito, exibe uma melhora significativa na coerência local e na construção de frases mais naturais. Adicionalmente, o tom e o vocabulário assemelham-se de forma mais evidente ao estilo pop característico da artista alvo. No entanto, mesmo com essas melhorias, o modelo ainda não foi capaz de realizar uma continuação que preservasse o sentido geral introduzido pela seed — que sugeria um sentimento de apego e exclusividade afetiva —, evidenciando limitações no controle semântico de longo prazo durante a geração textual.

5. CONCLUSÕES

Este estudo demonstrou a viabilidade do uso de modelos LSTM bidirecionais em nível de caractere para a geração de letras musicais estilisticamente coerentes com um artista específico. A combinação de um pré-treinamento em um corpus amplo com posterior fine-tuning utilizando dados exclusivos da artista Ariana Grande mostrou-se particularmente eficaz, resultando em melhorias notáveis na acurácia de predição de caracteres, na diversidade lexical e na redução de repetições de padrões linguísticos.

O fine-tuning destacou-se como um componente essencial do processo, elevando de forma significativa a qualidade das letras geradas em comparação com modelos treinados apenas com dados do artista ou com o modelo geral isolado. As letras produzidas pelo modelo ajustado apresentaram maior fluidez local e

refletiram mais claramente o estilo lírico característico da artista-alvo.

As diferentes estratégias de regularização por dropout aplicadas nos modelos iniciais não trouxeram ganhos consistentes. O modelo sem dropout, em muitos casos, obteve os melhores resultados, sugerindo que, neste cenário com dados relativamente limitados, o uso de dropout pode ter inibido a aprendizagem de padrões específicos e relevantes do corpus.

Apesar dos resultados promissores, algumas limitações persistem. A coerência semântica global das letras geradas, mesmo pelo modelo fine-tuned, ainda é um desafio. As frases podem ser localmente coerentes, mas a música como um todo pode não convergir para um tema central claro ou desenvolver uma narrativa consistente. Além disso, o contexto de entrada limitado a 100 caracteres pode restringir a capacidade do modelo de aprender dependências de muito longo alcance, que são importantes para a estrutura e o fluxo de uma letra de música completa.

Como próximos passos, propõe-se a exploração de arquiteturas mais modernas, como o CharacterBERT, uma variação do BERT que opera em nível de caractere e gera embeddings contextuais — ou seja, a representação de um caractere muda conforme seu uso na frase, o que pode enriquecer a modelagem de estilo e semântica em tarefas como geração de letras [17].

Outra possibilidade a ser explorada é a comparação entre a abordagem atual em nível de caractere e modelos que operam em nível de palavra. Essa alternativa pode trazer ganhos em coerência semântica, ainda que exija lidar com vocabulários maiores e o problema de palavras fora do vocabulário (OOV).

Além disso, aumentar a janela de contexto fornecida como entrada ao modelo pode ampliar sua capacidade de capturar dependências mais distantes no texto, contribuindo para letras mais estruturadas. Por fim, integrar informações estruturais da música — como padrões de rima, métrica ou contagem de sílabas — pode permitir um controle mais preciso da forma e estilo das composições geradas.

Em suma, este trabalho reforça o potencial do deep learning, em especial das LSTMs bidirecionais em nível de caractere, como ferramentas de apoio à composição musical. Os resultados obtidos fornecem uma base sólida para investigações futuras voltadas à geração criativa de texto musical com maior fluidez, personalização e consistência.

6. AGRADECIMENTOS

Gostaríamos de agradecer ao Prof. Dr. Wemerson D. Parreira pelas orientações e suporte durante o desenvolvimento deste trabalho na disciplina de Processamento de Linguagem Natural.

7. REFERÊNCIAS

[1] JOHNSON, G. O Processo Criativo de Composição de uma Nova Peça de Música. Horn Society Journal, [S. l.]: [s. n.], 2021. p. [s.p.]. Acesso em: 5 jun. 2025.

[2] KROL, A.; LIU, J.; JIN, X.; AGAPIE, E. Exploring the Needs of Practising Musicians in Co-Creative AI Through Co-Design. In: ACM CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (CHI), 2025, Yokohama, Japan. Anais... New York, NY: ACM, 2025. p. [s.p.]. Acesso em: 5 jun. 2025.

[3] NEWMAN, A.; JAIN, A.; LOUIE, S.; SCHEDL, M. End-to-End Human-AI Music Creation: Artists' Perceptions of AI Technology. Transactions of the International Society for Music Information Retrieval, London, v. 6, n. 1, p. 1-22, 2023. Acesso em: 5 jun. 2025.

[4] GERA, S. The Impact of Artificial Intelligence on Music Production: Creative Potential, Ethical Dilemmas, and the Future of the Industry. National High School Journal of Science, [S. l.], v. 1, p. 1, 2025. Acesso em: 5 jun. 2025.

[5] MICCHI, G.; CAÑAMERO, L.; MCKINNEY, M. I Keep Counting: An Experiment in Human/AI Co-creative Songwriting. Transactions of the ISMIR, London, v. 4, n. 1, p. 263-275, 2021. Acesso em: 5 jun. 2025.

[6] KIM, Y.; LEE, S.-J.; DONAHUE, C. AMUSE: Human-AI Collaborative Songwriting with Multimodal Inspirations. In: ACM CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (CHI), 2025, Yokohama, Japan. Anais... New York, NY: ACM, 2025. p. [s.p.]. Acesso em: 5 jun. 2025.

[7] LYRICSTUDIO. LyricStudio: AI-Powered Lyrics Platform. [S. l.]: Songwriter's Pad, [s.d.]. Disponível em: https://lyricstudio.net. Acesso em: 5 jun. 2025.

[8] SUNO AI. Suno AI. [S. l.]: Suno, Inc., 2023. Disponível em: https://suno.com. Acesso em: 5 jun. 2025.

[9] ILAKIYASELVAN, N.; MANDAL, S.; BHADRA, S.; DHANDAPANI, A.; VISWANATHAN, V. Lyrics generation using LSTM and RNN. In: MANDAL, J. K.; DE, D. (org.). Lecture Notes in Electrical Engineering. [S.l.]: Springer, 2023. p. 275–291. DOI: https://doi.org/10.1007/978-981-99-1051-9_24. Disponível em: https://www.researchgate.net/publication/371484570_Lyrics_Generation_Using_LSTM_and_RNN. Acesso em: 5 jun. 2025.

[10] GRAVES, A. Generating Sequences With Recurrent Neural Networks. [S. l.]: [s. n.], 2013. Disponível em: https://arxiv.org/abs/1308.0850. Acesso em: 5 jun. 2025.

[11] DIPIETRO, R.; HAGER, G. D. Deep learning: RNNs and LSTM. In: ZHOU, S. K.; RUECKERT, D.; FICHTINGER, G. (ed.). Handbook of Medical Image Computing and Computer Assisted Intervention. London; San Diego, CA: Academic Press, 2020. p. 503-519. Acesso em: 5 jun. 2025.

[12] GRAVES, A.; FERNÁNDEZ, S.; SCHMIDHUBER, J. Bidirectional LSTM networks for improved phoneme classification and recognition. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL NEURAL NETWORKS, 15., 2005, Warsaw, Poland. Proceedings... Berlin: Springer, 2005. p. 799-804. Acesso em: 5 jun. 2025.

[13] WANG, P.; QIAN, Y.; SOONG, F. K.; HE, L.; ZHAO, H. Learning distributed word representations for bidirectional LSTM recurrent neural network. In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, 2016, San Diego, CA. Proceedings... San Diego, CA: Association for Computational Linguistics, 2016. p. 527-533. Acesso em: 5 jun. 2025.

[14] ZAREMBA, W.; SUTSKEVER, I.; VINYALS, O. Recurrent Neural Network Regularization. [S. l.]: [s. n.], 2014. Disponível em: https://arxiv.org/abs/1409.2329. Acesso em: 20 maio 2024.

[15] KIM, Y.; JERNITE, Y.; SONTAG, D.; RUSH, A. M. Character-Aware Neural Language Models. [S. l.]: [s. n.], 2015. Disponível em: https://arxiv.org/abs/1508.06615. Acesso em: 5 jun. 2025.

[16] GURURANGAN, S.; MARASOVIĆ, A.; SWAYAMDICTA, S.; LO, K.; BELTAGY, I.; DOWNEY, D.; SMITH, N. A. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. [S. l.]: [s. n.], 2020. Disponível em: https://arxiv.org/abs/2004.10964. Acesso em: 5 jun. 2025.

[17] H. El Boukkouri, O. Ferret, T. Lavergne, H. Noji, P. Zweigenbaum, and J. Tsujii. *CharacterBERT: Reconciling ELMo and BERT for Word-Level Open-Vocabulary Representations From Characters*. In Proceedings of the 28th International Conference on Computational Linguistics (COLING), pp. 6903–6915, 2020. Acesso em: 5 jun. 2025.