INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS PICOS
ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

VICTOR GABRIEL PEREIRA DE SOUSA

MULTITASK LEARNING APLICADO À CLASSIFICAÇÃO DE UTILIDADE DE OPINIÕES

PICOS, PIAUÍ
2025

VICTOR GABRIEL PEREIRA DE SOUSA

MULTITASK LEARNING APLICADO À CLASSIFICAÇÃO DE UTILIDADE DE OPINIÕES

Trabalho de Conclusão de Curso (Artigo Científico) apresentado como exigência parcial para obtenção do diploma do Curso de Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Picos.

Orientador: Prof. Dr. Rogério Figueredo de Sousa

PICOS, PIAUÍ
2025

MULTITASK LEARNING APLICADO À CLASSIFICAÇÃO DE UTILIDADE DE OPINIÕES

Victor Gabriel Pereira de Sousa[1]
Rogério Figueredo de Sousa[2]

RESUMO

O estudo investiga a aplicação do *Multitask Learning* (MTL) na classificação de utilidade de opiniões, considerando a interação entre polaridade (positiva/negativa) e utilidade (útil/não útil) como tarefas complementares. A pesquisa visou desenvolver um modelo mais preciso e generalizável para análise de sentimentos e filtragem de conteúdo.

Inicialmente, o artigo revisa trabalhos relacionados, como o estudo de (SOUSA, 2023), que utiliza aprendizado de máquina para classificar opiniões úteis em português brasileiro, e a abordagem de (SATAPATHY; PARDESHI; CAMBRIA, 2022), que emprega MTL e BERT *embeddings* para análise de polaridade e subjetividade. A fundamentação teórica abrange temas como Processamento de Linguagem Natural (PLN), aprendizado multitarefa e classificação de utilidade de opiniões. A metodologia envolve a utilização do UTLCorpus, um conjunto de 2.881.589 avaliações de aplicativos Android e filmes, rotuladas automaticamente quanto à utilidade e polaridade. O estudo compara modelos de *Single-Task Learning* (STL) e MTL, utilizando redes neurais e BERT para classificar as opiniões e com isso foram alcançadas métricas como 0.64 para classificação de utilidade de opinião e 0.91 para classificação de polaridade usando o BERT.

Palavras-chaves: Multitask Learning, Processamento de Liguagem Natural, Utilidade de opinões.

ABSTRACT

The study investigates the application of Multitask Learning (MTL) in opinion helpfulness classification, considering the interaction between polarity (positive/negative) and helpfulness (helpful/not helpful) as complementary tasks. The research aims to develop a more precise and generalizable model for sentiment analysis and content filtering.

Initially, the paper reviews related works, such as the study conduded by (SOUSA, 2023), which applies machine learning to classify helpful opinions in Brazilian Portuguese, and the approach by (SATAPATHY; PARDESHI; CAMBRIA, 2022), which employs MTL and BERT embeddings for polarity and subjectivity analysis. The theoretical framework covers topics such as Natural Language Processing (NLP), multitask learning, and opinion helpfulness classification.

The methodology involves the use of UTLCorpus, a dataset containing 2,881,589 reviews of Android applications and movies, automatically labeled for helpfulness and polarity. The study compares Single-Task Learning (STL) and MTL models, utilizing neural networks and BERT to classify opinions and with that, metrics such as 0.64 for opinion utility classification and 0.91 for polarity classification using BERT were achieved.

[1] http://lattes.cnpq.br/0647540220826462. IFPI. capic.2022118tads0021@aluno.ifpi.edu.br.
[2] http://lattes.cnpq.br/9346694618502913. IFPI. rogerio.sousa@ifpi.edu.br.

Keywords: Multitask Learning, Natural Language Processing, Opinion helpfulness.

Data de Aprovação: 02 de Abril de 2025(Data de Apresentação do Artigo)

1 INTRODUÇÃO

A grande quantidade de opiniões disponíveis em ambientes digitais, especialmente em plataformas de avaliação e de *e-commerce*, tem incentivado o desenvolvimento de sistemas capazes de identificar automaticamente quais dessas opiniões são realmente úteis para os usuários. A tarefa de classificação de utilidade de opiniões visa separar informações relevantes daquelas irrelevantes, contribuindo para melhorar a experiência do usuário e a tomada de decisão em ambientes online.

E uma das áreas que visa resolver esse problema é o Processamento de Linguagem Natural (PLN) campo da inteligência artificial (IA) que busca desenvolver métodos para que máquinas compreendam, interpretem e gerem linguagem humana de forma eficiente (CASELI; NUNES, 2024). Essa área combina conceitos de linguística e aprendizado de máquina para lidar com diversas tarefas, como tradução automática, análise de sentimentos e classificação de textos. No contexto de plataformas digitais, o PLN desempenha um papel importante na organização e filtragem de grandes volumes de informações geradas pelos usuários.

Uma das aplicações relevantes do PLN é a classificação de utilidade de opiniões, que consiste em identificar quais comentários ou avaliações são mais úteis para outros usuários. Esse processo analisa aspectos como clareza, profundidade da informação e relevância do conteúdo, permitindo que as plataformas destaquem opiniões mais valiosas (SOUSA, 2023). Dessa forma, os usuários podem tomar decisões mais embasadas ao escolher um filme, um produto ou um serviço, tornando a experiência digital mais eficiente e satisfatória.

Dentre as técnicas utilizadas no aprendizado de máquina, existe a técnica de *Multitask Learning* (Aprendizado Multitarefa), que busca combinar duas ou mais tarefas de forma que elas possam compartilhar conhecimento durante o treinamento. A ideia central é que as tarefas envolvidas estejam relacionadas de alguma maneira, permitindo que o modelo aprenda representações mais generalizáveis e eficazes (WANG; ZHAI; AWADALLA, 2020). Com base nessa estratégia, este projeto tem como objetivo investigar o uso do *Multitask Learning* para aprimorar a capacidade de generalização de modelos na classificação da utilidade de opiniões. A proposta é explorar a sinergia entre duas tarefas relacionadas e complementares: a classificação de polaridade de opiniões (positivo ou negativo) e a classificação de utilidade de opiniões (útil ou não-útil). Ambas as tarefas são binárias e compartilham uma relação intrínseca, uma vez que a polaridade de uma opinião pode influenciar sua percepção de utilidade, e vice-versa.

Ao treinar um modelo para aprender essas tarefas de forma conjunta, espera-se que ele capture padrões mais robustos e generalize melhor para novos dados, reduzindo o risco de *overfitting* e melhorando a eficiência do aprendizado (CARUANA, 1997; RUDER, 2017; ZHANG; YANG, 2021).

A tarefa de classificação de utilidade de opinião é particularmente relevante em cenários onde grandes volumes de opiniões são gerados, como em plataformas de streaming de filmes e lojas de aplicativos. A capacidade de identificar opiniões úteis e relevantes pode melhorar significativamente a experiência do usuário, auxiliando-o na tomada de decisões mais informadas, como a escolha de um filme ou aplicativo. Além disso, o *Multitask Learning* tem se mostrado eficaz em tarefas de processamento de linguagem natural (PLN), onde a aprendizagem conjunta de tarefas relacionadas pode levar a uma melhor compreensão de nuances contextuais e a uma representação mais rica dos dados textuais (LIU *et al.*, 2019).

Este estudo busca contribuir para a literatura ao demonstrar como o *Multitask Learning* pode ser aplicada na classificação de utilidade de opiniões, explorando a interação entre polaridade e utilidade como tarefas complementares. A meta é desenvolver um modelo que não apenas identifique opiniões úteis com maior precisão, mas que também seja capaz de generalizar para diferentes contextos e domínios.

O restante deste trabalho está organizado da seguinte forma: na Seção 2, são apresentados os trabalhos relacionados, abordando pesquisas ligadas ao tema de utilidade de opinião e ao uso de Multitask Learning na análise de sentimento. A Seção 3 (Referencial Teórico) fornece uma visão geral dos conceitos fundamentais para o entendimento deste artigo. Em seguida, a Seção 4 descreve a metodologia utilizada na pesquisa. Na Seção 5, são apresentados os resultados obtidos nos experimentos, incluindo comparações com modelos da literatura. Por fim, a Seção 6 encerra o trabalho com as conclusões e considerações finais.

2 TRABALHOS RELACIONADOS

2.1. CLASSIFICAÇÃO DA UTILIDADE DE OPINIÕES EM PORTUGUÊS BRASILEIRO

O trabalho de (SOUSA, 2023) investiga a tarefa de classificação automática da utilidade de opiniões no contexto do português brasileiro, aplicando aprendizado de máquina para identificar quais comentários são úteis para os usuários. A pesquisa explora fatores linguísticos e contextuais, criando um corpus multidomínio que engloba opiniões sobre aplicativos de smartphone e filmes. A abordagem do autor foca em métodos de classificação, utilizando um método baseado em grafos para classificar a utilidade das opiniões. Além disso, destaca-se a proposta de um *benchmark* para

a tarefa, com resultados relevantes na classificação de opiniões sobre aplicativos. Embora a tese de Sousa tenha contribuído significativamente para o campo, ela se limita ao uso de técnicas de classificação tradicionais, abrindo espaço para a exploração de abordagens como o multitask learning, que pode melhorar o desempenho ao lidar com tarefas inter-relacionadas, como classificação de utilidade e análise de sentimentos simultaneamente.

2.2. DETECÇÃO DE POLARIDADE E SUBJETIVIDADE COM APRENDIZADO MULTITAREFA E EMBEDDINGS BERT

O trabalho de (SATAPATHY; PARDESHI; CAMBRIA, 2022) propõe um modelo de aprendizado multitarefa (MTL) para análise de sentimentos, combinando a detecção de polaridade (positiva ou negativa) e a detecção de subjetividade (subjetivo ou objetivo). Utilizando *embeddings* BERT e uma Rede Neural Tensorial (NTN) para compartilhar conhecimento entre as tarefas, os autores demonstram que essa abordagem melhora a precisão dos modelos em comparação com métodos tradicionais. A técnica explora a interdependência entre as tarefas, permitindo que informações de uma auxiliem na previsão da outra, o que reduz o risco de *overfitting* e melhora a generalização do modelo.

Os experimentos realizados mostram que a abordagem MTL supera os modelos baseados em aprendizado de tarefa única, destacando o potencial da técnica para aprimorar a extração de *insights* de textos, especialmente em aplicações como pesquisas de mercado. O artigo também discute diferentes estratégias de compartilhamento de parâmetros em MTL, optando pelo compartilhamento rígido (hard parameter sharing) para melhorar a eficiência e o desempenho da rede.

3 REFERENCIAL TEÓRICO

3.1. PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

O Processamento de Linguagem Natural (PLN) é uma área central da inteligência artificial que busca a interação eficaz entre humanos e máquinas por meio da linguagem natural. O PLN desempenha um papel na interação entre humanos e máquinas, possibilitando avanços na compreensão automática da linguagem e na automação de diversas tarefas que envolvem texto (ANCHIÊTA *et al.*, 2021). Um exemplo de modelos que são voltados ao PLN são o BERT (DEVLIN, 2018) e GPT (RADFORD; WU, 2019), que trouxeram avanços significativos na compreensão semântica e sintática de textos, permitindo a construção de sistemas que captam melhor as nuances de linguagem. Aplicações práticas de PLN incluem a análise de opiniões em plataformas de e-commerce, onde a compreensão automática de textos é crucial para fornecer

recomendações e avaliações mais precisas (CASELI; NUNES, 2024). No contexto de análise de opiniões, a habilidade de um modelo em captar a subjetividade do texto é um desafio que o PLN busca resolver. Como destaca (CAMBRIA *et al.*, 2017), a análise de sentimentos se tornou uma aplicação de grande valor comercial e acadêmico, e as técnicas de PLN são fundamentais para o sucesso dessas análises em escala.

3.2. MULTITASK LEARNING

Multitask Learning (MTL) é uma abordagem no aprendizado de máquina em que múltiplas tarefas são aprendidas simultaneamente com o objetivo de melhorar a generalização dos modelos ao explorar informações compartilhadas entre tarefas. (CARUANA, 1997) foi um dos primeiros a definir MTL como uma técnica de aprendizagem em que as tarefas relacionadas são usadas para melhorar a performance geral do modelo. A principal vantagem do MTL, conforme (RUDER, 2017), é que ao compartilhar representações entre tarefas correlatas, o modelo pode aprender de maneira mais eficiente e reduzir o risco de *overfitting*. No contexto de PLN, MTL tem sido amplamente aplicado para tarefas de classificação de texto, incluindo a classificação de polaridade de opiniões e a classificação de utilidade de opiniões. (ZHANG; YANG, 2021) exploram como o MTL pode ser usado em PLN para melhorar a precisão de modelos ao realizar múltiplas tarefas de processamento de texto simultaneamente. Esse processo de compartilhamento de informações auxilia na captura de padrões semânticos complexos que são relevantes para diferentes tarefas, resultando em um modelo mais robusto e eficiente.É importante também citar alguns dos tipos de MLT, como o *Hard Parameter Sharing* onde os parâmetros iniciais da rede são compartilhados entre todas as tarefas, e as últimas camadas são específicas para cada tarefa, e o *Soft Parameter Sharing* que em vez de compartilhar os mesmos parâmetros, cada tarefa tem sua própria rede neural, mas as redes compartilham informações por meio de regularização.

3.3. UTILIDADE DE OPINIÃO

A classificação de utilidade de opiniões é uma tarefa complexa e multifacetada, especialmente no contexto de plataformas de *e-commerce*, onde grandes volumes de opiniões são gerados diariamente. (KIM *et al.*, 2006) definem a utilidade de opinião como o valor de uma opinião com base na sua capacidade de ajudar outros usuários a tomar decisões informadas. Estudos sobre a identificação de opiniões úteis frequentemente utilizam métodos baseados em aprendizado de máquina para automatizar a classificação dessas opiniões, levando em consideração fatores como clareza, detalhamento e relevância. No entanto, como apontado por (LIU, 2022), a utilidade de uma opinião é subjetiva e pode variar significativamente dependendo do contexto e das necessidades do usuário. Isso torna a tarefa de identificar automaticamente opiniões

úteis um desafio considerável para os modelos de aprendizado de máquina. Como observado por (MCAULEY; LESKOVEC, 2013), que argumentam que tarefas como análise de sentimentos e utilidade de opiniões estão frequentemente inter-relacionadas, o que pode ser aproveitado por abordagens multitarefa.

3.4. REDES NEURAIS CONVOLUCIONAIS (CNNS')

Originalmente desenvolvidas para visão computacional, as Redes Neurais Convolucionais (CNNs) mostraram-se eficazes também em diversas tarefas de PLN, conforme demonstrado por (KIM, 2014). Nesse contexto, as CNNs operam sobre representações vetoriais de textos, aplicando filtros convolucionais capazes de extrair padrões locais, como n-gramas relevantes, que contribuem para a construção de representações discriminativas dos textos.

O principal diferencial das CNNs no processamento de linguagem é sua capacidade de capturar padrões estruturais fixos e localmente relevantes, mesmo com representações densas e de curta duração. Apesar de sua limitação em capturar dependências de longo prazo, as CNNs têm sido amplamente utilizadas em sistemas de classificação de textos, incluindo a avaliação de opiniões, onde o reconhecimento de estruturas linguísticas locais pode ser decisivo.

3.5. O BERT (BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS)

proposto por (DEVLIN, 2018), revolucionou as abordagens em Processamento de Linguagem Natural (PLN) ao introduzir um modelo pré-treinado de forma bidirecional profunda, baseado na arquitetura Transformer. Diferente de modelos anteriores, o BERT realiza treinamento com máscaras em palavras aleatórias (*masked language modeling*) e previsão de próxima sentença (*next sentence prediction*), permitindo que aprenda representações contextuais mais ricas para palavras em diferentes posições.

A bidirecionalidade do BERT possibilita que ele capture dependências linguísticas de longo alcance e compreenda melhor o contexto em que as palavras aparecem, o que é especialmente útil em tarefas como classificação de texto, análise de sentimentos e extração de informações. Após o pré-treinamento, o BERT pode ser ajustado (*fine-tuned*) para tarefas específicas com apenas algumas camadas adicionais, mantendo grande parte do conhecimento linguístico aprendido em grandes corpora.

4 METODOLOGIA

Com a finalidade de atingir o objetivo principal, este artigo foi dividido em quatro etapas, inicialmente será apresentado o dataset que será utilizado e em seguida serão

destacadas as etapas:

4.1. UTLCORPUS

O UTLCorpus é um córpus de opiniões em português do Brasil, focado em avaliações de aplicativos para Android e filmes, desenvolvido para suportar pesquisas de classificação de utilidade e polaridade de opiniões online. Criado como parte de um trabalho de doutorado, o córpus visa suprir a carência de bases de dados em português brasileiro que contenham anotações sobre a utilidade de avaliações. As avaliações foram extraídas de uma rede social brasileira de filmes e da Google Play Store, sendo pré-processadas e anonimizadas. Esse córpus é um extenso conjunto de dados em português do Brasil, composto por 2.881.589 avaliações, sendo 1.839.851 de filmes e 1.041.738 de aplicativos Android. O córpus foi criado com o objetivo de apoiar pesquisas sobre a predição de utilidade (*helpfulness*) e a classificação de polaridade em avaliações online. As avaliações foram coletadas por meio de *web crawlers*, escolhendo esses domínios pela sua popularidade, alta quantidade de dados e a presença de um contador público de *likes*, que permitiu inferir rótulos de utilidade (SOUSA; BRUM; NUNES, 2019).

4.2. ETAPAS DO TRABALHO

1. **Definição de tarefas correlatas:**

   Foram avaliadas várias tarefas que possuem uma forte correlação com a classificação de utilidade de opiniões, como a classificação de polaridade, a detecção de sentimentos, a identificação de tópicos e a extração de aspectos. Contudo dadas as limitações em relação à anotação para cada uma das tarefas, a que mais se encaixava com o tempo disponível para realização do trabalho foi a de classificação de polaridade, e além disso, já possui alguma referência para comparação que é o trabalho do UTLCorpus.

2. **Implementação e/ou replicação dos modelos individuais e realização de experimentos com MTL:**

   Nesta fase, foi realizada a implementação de modelos de aprendizado de máquina para a classificação de opiniões, tanto de forma individual quanto utilizando *multitask learning* (MTL). Foram explorados duas abordagens distintas:

   - **Modelo Base** [3] (Classificação de Polaridade Simples): Para estabelecer um modelo de referência para a tarefa de classificacão de polaridade, foi implementado um classificador simples de polaridade de opinião. Esse

[3] <https://github.com/victgab20/polaridade-analise-utl_corpus>

modelo consiste em uma camada linear seguida de uma função de ativação sigmoide, responsável por prever se uma opinião é positiva ou negativa.

- **Modelos de Aprendizado Multitarefa** (MTL): Além do modelo simples, foram implementados dois modelos utilizando *multitask learning*, nos quais a rede realiza simultaneamente a classificação de polaridade da opinião (negativa ou positiva) e a classificação de utilidade da opinião. As arquiteturas utilizadas foram:
  - **MultiTaskBertModel**: Baseado na arquitetura BERT (*Bidirectional Encoder Representations from Transformers* [4]), este modelo utiliza a representação contextualizada de texto fornecida pelo BERT para gerar uma única representação para cada entrada. Essa representação é então processada por duas camadas lineares separadas: uma dedicada à predição da polaridade da opinião e outra à predição da utilidade da opinião. Ambas as saídas são obtidas a partir do *pooler output* do BERT, que corresponde à representação do token [CLS], e passam por camadas de dropout para regularização.
  - **CNN Multitask** [5]: Inspirado em abordagens baseadas em convolução para modelagem de texto, esse modelo utiliza uma rede convolucional para extrair padrões locais da sequência de palavras. O modelo incorpora camadas de convolução 1D aplicadas a embeddings pré-treinados, seguidas de operações de *max pooling*, resultando em uma representação compacta do texto. Essa representação final é então processada por uma rede totalmente conectada, que se ramifica em duas cabeças de saída: uma para a predição da polaridade da opinião e outra para a predição da utilidade da opinião. As saídas finais são obtidas por camadas lineares com ativação sigmoide.

A abordagem multitarefa explora a relação entre as tarefas de classificação de polaridade e utilidade, permitindo que o modelo compartilhe informações entre as duas tarefas e, potencialmente, aprenda representações mais robustas do texto.

5 RESULTADOS

Nessa seção vamos apresentar os resultados alcançados neste trabalho e uma comparação com os dados presentes na literatura.

[4] <https://github.com/victgab20/bert_model>
[5] <https://github.com/victgab20/cnn_teste>

5.1. CLASSIFICAÇÃO DE UTILIDADE

Neste tópico, exploramos como a classificação de utilidade de opinião pode ser abordada utilizando o paradigma *Single-Task Learning*, ou seja, treinando um modelo específico para essa tarefa, sem o auxílio de outras tarefas relacionadas.

Para isso, analisamos diversos trabalhos que abordaram o problema da classificação de utilidade sob a perspectiva *Single-Task Learning*, descrevendo seus métodos, técnicas, resultados e principais contribuições. Aqui está uma tabela listando alguns deles:

Tabela 1 – Trabalhos relacionados à avaliação da utilidade de opinião.

Trabalhos	**Abordagem**	**Método**	**Resultado**
(KIM *et al.*, 2006)	Regressão e Raking	SVT	0,656 e 0,604 de *spearman*
(ZHANG; VARADARAJAN, 2006)	Regressão	SVR e SLR	0,056 e 0,082 de MSE livros e filmes, respectivamente
(GHOSE; IPEIROTIS, 2007)	Classificação	Regressão Logística	F1: 0,85
(GHOSE; IPEIROTIS, 2010)	Classificação	Random Forest	F1: 89,04 em produtos; ROC: 0,94
(LIU *et al.*, 2008)	Regressão	RBF	MSE: 0,033
(SOUSA; PARDO, 2022)	Classificação	CNN	F1: apps : 0,91; filmes: 0.74

Fonte: Adaptado de Sousa (2023).

A Tabela 1 apresenta um panorama dos principais trabalhos relacionados à avaliação da utilidade de opiniões, destacando suas abordagens, métodos empregados e os principais resultados obtidos. Os estudos variam entre técnicas de regressão, classificação e ranqueamento, utilizando diferentes algoritmos como Support Vector Regression (SVR), Regressão Logística, Random Forest, Redes Neurais Radiais (RBF) e Redes Neurais Convolucionais (CNN).

Dentre os trabalhos analisados, destaca-se (GHOSE; IPEIROTIS, 2010), que obteve um valor elevado de F1-score (89,04) na classificação de opiniões sobre produtos, além de uma AUC-ROC de 0,94. Já (SOUSA; PARDO, 2022) aplica CNNs para classificar opiniões de aplicativos e filmes, alcançando F1-scores de 0,91 e 0,74, respectivamente. Esses resultados reforçam a tendência recente de adoção de modelos baseados em aprendizado profundo, capazes de extrair características relevantes diretamente do texto.

Por outro lado, abordagens mais tradicionais como Regressão Logística e SVR continuam apresentando desempenho competitivo, como observado nos trabalhos de (KIM *et al.*, 2006) e (GHOSE; IPEIROTIS, 2007), com bons indicadores de correlação e precisão. A diversidade de métodos e métricas evidencia a complexidade do problema e a ausência de uma abordagem dominante, o que justifica a investigação proposta neste trabalho.

Neste trabalho usamos para comparação os resultados obtidos no trabalho de

(SOUSA; BRUM; NUNES, 2019) onde foram obtidas as seguintes métricas para á classificação de polaridade:

Tabela 2 – Resultados de detecção de utilidade

Classifier	**Movie Review**			**App Review**		
	F1-Help	F1-No-Help	F1-Measure	F1-Help	F1-No-Help	F1-Measure
Baseline	0.4499	0.6493	0.5496	0.5896	0.6617	0.6256
Linear SVM	0.6341	0.6039	0.6189	0.7115	0.6436	0.6775
MLP	0.6387	0.6118	**0.6252**	0.7082	0.6516	0.6799
Random Forest	0.6220	0.5920	0.6072	0.7267	0.7182	**0.7224**

Fonte: Adaptado de Sousa (2023).

A Tabela 2 apresenta os resultados obtidos no trabalho de (SOUSA; BRUM; NUNES, 2019) para a tarefa de classificação de utilidade de opiniões, tanto em críticas de filmes (Movie Review) quanto em avaliações de aplicativos (App Review). As métricas reportadas incluem os valores de F1-score para as classes "Help"(útil), "No-Help"(não útil) e a média entre ambas (F1-Measure), permitindo uma análise mais equilibrada entre precisão e revocação nos diferentes cenários.

Na base de críticas de filmes, os melhores resultados foram alcançados com o classificador MLP (Perceptron Multicamadas), com F1-Measure de 0,6252, enquanto a Linear SVM e a Random Forest apresentaram desempenhos próximos, com F1-scores médios de 0,6189 e 0,6072, respectivamente. Para as avaliações de aplicativos, o melhor desempenho foi obtido pela Random Forest, com um F1-Measure de 0,7224, evidenciando sua robustez para esse domínio específico.

É importante observar que, em ambos os domínios, os classificadores supervisionados superaram significativamente o desempenho da linha de base (baseline), o que demonstra a viabilidade de modelos tradicionais de aprendizado de máquina para a tarefa de identificação de opiniões úteis. Esses resultados são utilizados neste trabalho como referência comparativa para avaliar a eficácia do modelo baseado em aprendizado multitarefa, proposto como alternativa aos métodos tradicionais.

O modelo criado para a tarefa de classificação de polaridade utilizou um total de 15 épocas e taxa de aprendizado igual a 0.001. Quanto ao tamanho de *batch* durante o treinamento de modelos para otimizar o uso de memória e acelerar a convergência, ao invés de processar todas as 1.188.408 amostras de uma vez, os dados são divididos em partes menores, com *batches* de 1500 amostras que são passadas para o modelo.

A definição da polaridade de referência dos comentários do dataset foi realizada por meio da conversão da coluna *stars* para valores numéricos, seguida da remoção de avaliações com nota igual a 3, pois representam uma zona neutra que pode introduzir ruído na classificação. A variável-alvo (label) foi então estabelecida, classificando avaliações com nota superior a 3 como positivas (1) e inferiores a 3 como negativas

(0). Em seguida, foram selecionadas apenas as colunas relevantes para o modelo, e eventuais valores ausentes foram preenchidos com zero para evitar inconsistências durante o treinamento.

Para representar os textos numericamente, utilizou-se o *TfidfVectorizer*[6], que converte as avaliações em vetores baseados na frequência inversa do termo (TF-IDF), limitando o vocabulário a 5.000 termos para otimizar o uso de memória. Além disso, os dados são processados em mini-batches de 500 amostras, permitindo maior eficiência computacional ao reduzir a carga de memória durante o treinamento do modelo.

Os resultados obtidos ao longo das épocas em ambos os dados de treinamento filmes e apps são apresentados na Tabela 3.

Tabela 3 – Resultados do modelo de polaridade

Modelo	**Filmes**				**Apps**			
	Acurácia	Precisão	Recall	F1 Score	Acurácia	Precisão	Recall	F1 Score
Rede Neural	0.7619	0.7724	0.9440	0.8496	0.9125	0.9301	0.9586	0.9441

Fonte: própria do autor

A Tabela 3 apresenta os resultados obtidos pelo modelo de classificação de polaridade em diferentes domínios de dados — críticas de filmes e avaliações de aplicativos. Os resultados correspondem às métricas de avaliação aferidas nos dados de teste após o treinamento do modelo, abrangendo acurácia, precisão, recall e F1-score, que permitem uma análise abrangente do desempenho.

Para o conjunto de dados de filmes, o modelo baseado em rede neural obteve uma acurácia de 76,19

No conjunto de dados de aplicativos, os resultados foram ainda mais expressivos, com acurácia de 91,25

5.2. CLASSIFICAÇÃO USANDO *MULTITASK*

A utilização de *multitask learning* parte da premissa de que tarefas relacionadas podem compartilhar representações internas e, assim, melhorar o desempenho geral do modelo.

Para este estudo, foram implementados dois modelos multitarefa um baseado no BERT adaptado para a classificação binária que foi treinado com 5 épocas e com um *learning rate* de 0.00002. A ideia principal foi aproveitar a mesma representação extraída do BERT para realizar duas tarefas distintas: polaridade (se a opinião expressa é positiva ou negativa) e utilidade (se a opinião é considerada útil ou não). E outro

[6] <https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html>

utilizamos uma *Convolucional Neural Network* (CNN), e em seu treinamento foi utilizado 5 épocas e um *learning rate* de 0.0001 e com embeddings pré-treinados do GloVe 300D, permitindo representar palavras como vetores densos. Para ambos os modelos mantivemos a divisão do córpus. As Tabelas 4 e 5 apresentam os resultados atigindos utilizando a parte de apps e filmes respectivamente :

Tabela 4 – Métricas de desempenho para as tarefas de classificação (Apps)

Modelo	**Helpfulness**			**Polarity**		
	Precisão	Recall	F1 Score	Precisão	Recall	F1 Score
Bert	0.61	0.59	0.64	0.78	0.77	**0.78**
CNN	0.75	0.70	**0.72**	0.72	0.86	**0.78**

Fonte: própria do autor

Tabela 5 – Métricas de desempenho para as tarefas de classificação (Filmes)

Modelo	**Helpfulness**			**Polarity**		
	Precisão	Recall	F1 Score	Precisão	Recall	F1 Score
Bert	0.66	0.67	**0.64**	0.87	0.82	**0.91**
CNN	0.58	0.68	**0.63**	0.74	0.97	0.84

Fonte: própria do autor

As Tabelas 4 e 5 apresentam os resultados comparativos de desempenho dos modelos BERT e CNN nas tarefas de classificação de utilidade e classificação de polaridade, considerando dois domínios distintos: avaliações de aplicativos (apps) e críticas de filmes (filmes). As métricas utilizadas foram precisão, revocação (recall) e F1-score, permitindo uma avaliação mais completa do desempenho dos modelos em ambas as tarefas.

Na Tabela 4, observam-se os resultados sobre o conjunto de dados de aplicativos. Para a tarefa de utilidade, o modelo CNN obteve melhor desempenho com F1-score de 0,72, superando o BERT (0,64), indicando maior equilíbrio entre precisão e recall nesse cenário. Já na tarefa de polaridade, ambos os modelos apresentaram desempenho semelhante, com F1-score de 0,78, o que sugere que tanto o BERT quanto a CNN conseguem capturar bem as nuances semânticas desse domínio.

Na Tabela 5, referente ao conjunto de críticas de filmes, o modelo BERT demonstrou superioridade na tarefa de polaridade, com F1-score de 0,91, contra 0,84 da CNN. No entanto, na tarefa de utilidade, os resultados foram mais próximos: o BERT obteve F1-score de 0,64 e a CNN de 0,63. Esses resultados mostram que, embora o BERT tenha maior capacidade de generalização na tarefa de polaridade, a CNN se mantém competitiva na classificação de utilidade.

De forma geral, os resultados sugerem que a escolha do modelo pode depender da tarefa específica e do domínio dos dados. Enquanto o BERT tende a ter melhor desempenho em polaridade, a CNN se mostra eficaz na classificação de utilidade, principalmente em avaliações de aplicativos. Essa análise comparativa reforça a motivação do presente trabalho em propor e avaliar uma abordagem baseada em aprendizado multitarefa, buscando integrar o melhor das duas tarefas em um único modelo compartilhado.

5.3. COMPARAÇÃO DE RESULTADOS

Inicialmente, são apresentadas as comparações de resultados relacionados à tarefa de classificação de polaridade. As Figuras 1 e 2 apresentam gráficos que resumem os resultados das métricas comparativas entre os modelos de *Multitask Learning* e o modelo de classificação de polaridade.

A Figura 1 apresenta um gráfico comparativo da tarefa de classificação de polaridade utilizando *multitask learning*, treinado com a seção de apps do córpus. As métricas de polaridade foram avaliadas no modelo implementado, mantendo a mesma divisão de dados. Já a Figura 2 mostra a comparação da mesma arquitetura, porém treinada com a seção de filmes do córpus, permitindo analisar o impacto da variação do domínio nos resultados.

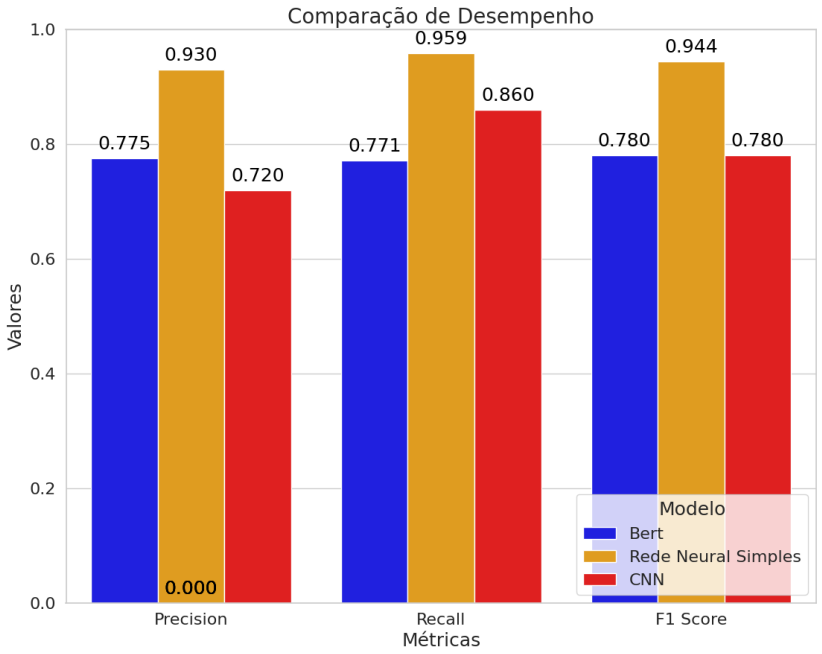

Figura 1 – Métricas de Polaridade Apps

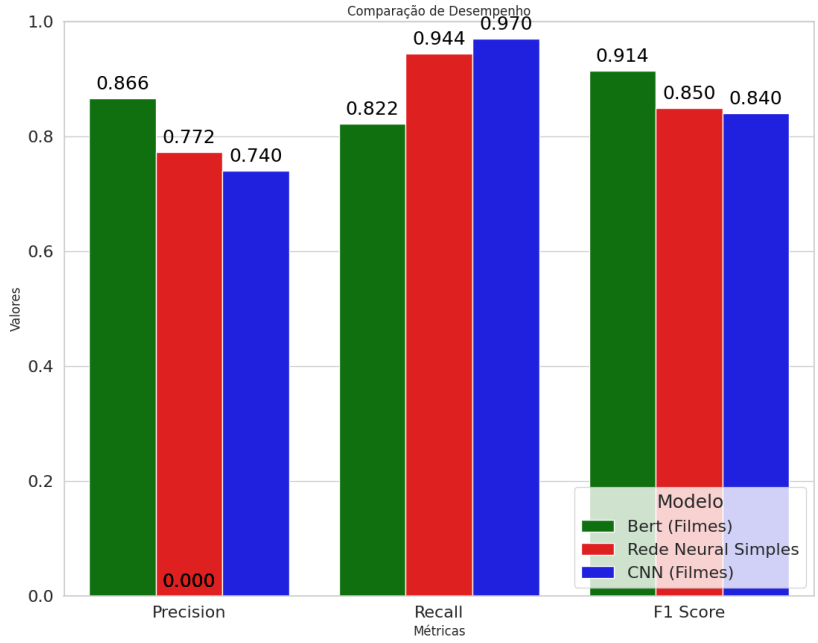

Figura 2 – Métricas de Polaridade Filmes

A Figura 2 apresenta um comparativo de desempenho entre os modelos Bert, Rede Neural Simples e CNN na tarefa de classificação de polaridade, treinados exclusivamente com dados da seção de filmes do córpus. As métricas avaliadas incluem Precisão, Recall e F1 Score. Observa-se que o modelo de Rede Neural Simples apresentou desempenho superior nas três métricas, com destaque para o Recall (0.959) e o F1 Score (0.944), indicando boa capacidade de identificar corretamente as classes e manter equilíbrio entre precisão e abrangência. O modelo Bert apresentou resultados mais modestos e consistentes entre as métricas (em torno de 0.77 a 0.78), enquanto a CNN teve desempenho inferior, especialmente na métrica de Precisão, com valor de apenas 0.000, sugerindo falhas na sua capacidade de prever corretamente exemplos positivos.

A Figura 3 apresenta a comparação da tarefa de classificação de utilidade utilizando Multitask Learning em relação aos modelos listados na Tabela 2, considerando a métrica F1 para os dados de aplicativos e filmes.

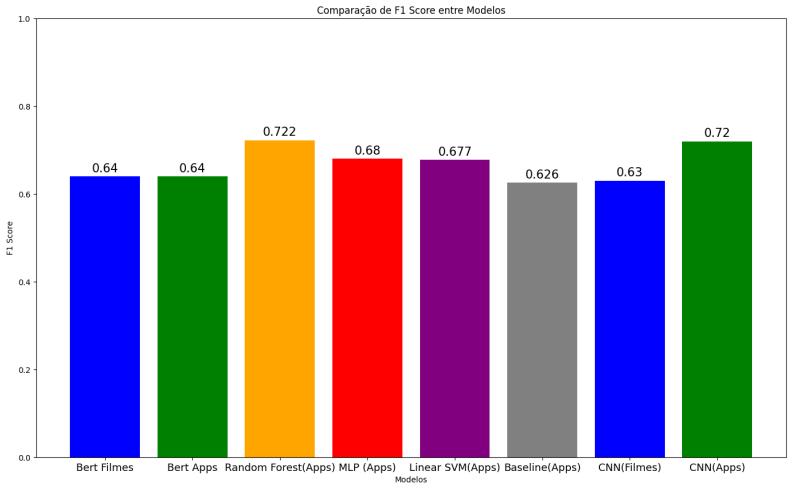

Figura 3 – Métricas de Utilidade

A Figura 3 apresenta a comparação de desempenho dos modelos na tarefa de classificação de utilidade utilizando a abordagem de Multitask Learning, com base na métrica de F1 Score. A análise contempla tanto os dados da seção de apps quanto de filmes, permitindo avaliar o impacto do domínio nos resultados. O modelo Random Forest treinado com dados de apps obteve o melhor desempenho geral, com F1 Score de 0.722, seguido de perto pela CNN (Apps) com 0.720. Em seguida, os modelos MLP (Apps) e Linear SVM (Apps) também apresentaram desempenhos competitivos, com 0.680 e 0.677, respectivamente. O modelo Bert, independentemente do domínio (filmes ou apps), mostrou desempenho inferior (ambos com 0.640), indicando que, nesse cenário, os modelos tradicionais superaram o modelo pré-treinado. O resultado mais fraco foi da abordagem Baseline (Apps) com 0.626, sugerindo a eficácia do fine-tuning e da modelagem supervisionada em tarefas de classificação mais complexas.

6 CONCLUSÃO OU CONSIDERAÇÕES FINAIS

Neste estudo, exploramos diferentes abordagens para a classificação de polaridade e utilidade de opiniões em português brasileiro, comparando métodos baseados em aprendizado de máquina tradicional e modelos baseados em aprendizado multitarefa (MTL). Utilizando o UTLCorpus, conduzimos experimentos com modelos individuais e multitarefa, avaliando o impacto da abordagem MTL no desempenho das tarefas.

Os resultados mostram que o uso de aprendizado multitarefa pode trazer benefícios consideráveis para a classificação de polaridade e utilidade. Em particular, observamos que o modelo baseado em BERT apresentou um desempenho não satisfatório tendo em vista o que se espera de um modelo tão robusto. Já a CNN, embora tenha apresentado bons resultados na tarefa de polaridade, teve um desempenho inferior na classificação de utilidade.

A comparação com trabalhos anteriores indica que, embora modelos tradicionais ainda apresentem desempenho competitivo, abordagens baseadas em MTL e redes neurais oferecem vantagens ao permitir o compartilhamento de informações entre tarefas correlatas. Isso sugere que o uso de aprendizado multitarefa pode ser um caminho promissor para aprimorar sistemas de recomendação e análise de opiniões em larga escala.

Dessa forma, este trabalho contribui para a literatura ao fornecer um *benchmark* para futuras pesquisas nessa área de classificação de utilidade de opinião.

Para trabalhos futuros, diversas abordagens podem ser exploradas para aprimorar o aprendizado multitarefa na classificação de opiniões. Métodos avançados, como *soft parameter sharing*, aprendizado adversarial e modelos hierárquicos, podem ser investigados para melhorar a generalização do modelo. Além disso, novas tarefas correlatas, como detecção de aspecto.

REFERÊNCIAS

ANCHIÊTA, R. *et al.* Pln: Das técnicas tradicionais aos modelos de deep learning. **Sociedade Brasileira de Computação**, 2021. Citado na página 5.

CAMBRIA, E. *et al.* Affective computing and sentiment analysis. **A practical guide to sentiment analysis**, Springer, p. 1–10, 2017. Citado na página 6.

CARUANA, R. Multitask learning. **Machine learning**, Springer, v. 28, p. 41–75, 1997. Citado 2 vezes nas páginas 4 e 6.

CASELI, H. M.; NUNES, M. G. V. (Ed.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 3. ed. BPLN, 2024. ISBN 978-65-01-20581-6. Disponível em: <https://brasileiraspln.com/livro-pln/3a-edicao/>. Citado 2 vezes nas páginas 3 e 6.

DEVLIN, J. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018. Citado 2 vezes nas páginas 5 e 7.

GHOSE, A.; IPEIROTIS, P. G. Designing novel review ranking systems: predicting the usefulness and impact of reviews. In: **Proceedings of the ninth international conference on Electronic commerce**. [S.l.: s.n.], 2007. p. 303–310. Citado na página 10.

GHOSE, A.; IPEIROTIS, P. G. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. **IEEE transactions on knowledge and data engineering**, IEEE, v. 23, n. 10, p. 1498–1512, 2010. Citado na página 10.

KIM, S.-M. *et al.* Automatically assessing review helpfulness. In: **Proceedings of the 2006 Conference on empirical methods in natural language processing**. [S.l.: s.n.], 2006. p. 423–430. Citado 2 vezes nas páginas 6 e 10.

KIM, Y. **Convolutional Neural Networks for Sentence Classification**. 2014. Disponível em: <https://arxiv.org/abs/1408.5882>. Citado na página 7.

LIU, B. **Sentiment analysis and opinion mining**. [S.l.]: Springer Nature, 2022. Citado na página 6.

LIU, X. *et al.* Multi-task deep neural networks for natural language understanding. **arXiv preprint arXiv:1901.11504**, 2019. Citado na página 4.

LIU, Y. *et al.* Modeling and predicting the helpfulness of online reviews. In: IEEE. **2008 Eighth IEEE international conference on data mining**. [S.l.], 2008. p. 443–452. Citado na página 10.

MCAULEY, J. J.; LESKOVEC, J. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In: **Proceedings of the 22nd international conference on World Wide Web**. [S.l.: s.n.], 2013. p. 897–908. Citado na página 7.

RADFORD, A.; WU, J. Rewon child, david luan, dario amodei, and ilya sutskever. 2019. **Language models are unsupervised multitask learners. OpenAI blog**, v. 1, n. 8, p. 9, 2019. Citado na página 5.

RUDER, S. An overview of multi-task learning in deep neural networks. **arXiv preprint arXiv:1706.05098**, 2017. Citado 2 vezes nas páginas 4 e 6.

SATAPATHY, R.; PARDESHI, S. R.; CAMBRIA, E. Polarity and subjectivity detection with multitask learning and bert embedding. **Future Internet**, MDPI, v. 14, n. 7, p. 191, 2022. Citado 2 vezes nas páginas 2 e 5.

SOUSA, R.; PARDO, T. Evaluating content features and classification methods for helpfulness prediction of online reviews: Establishing a benchmark for portuguese. In: **Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis**. [S.l.: s.n.], 2022. p. 204–213. Citado na página 10.

SOUSA, R. F. d. **Classificação da utilidade de opiniões em português brasileiro**. Tese (Doutorado) — Universidade de São Paulo, 2023. Citado 5 vezes nas páginas 2, 3, 4, 10 e 11.

SOUSA, R. F. d.; BRUM, H. B.; NUNES, M. d. G. V. A bunch of helpfulness and sentiment corpora in brazilian portuguese. In: **Proceedings**. [S.l.: s.n.], 2019. Citado 2 vezes nas páginas 8 e 11.

WANG, Y.; ZHAI, C.; AWADALLA, H. H. **Multi-task Learning for Multilingual Neural Machine Translation**. 2020. Disponível em: <https://arxiv.org/abs/2010.02523>. Citado na página 3.

ZHANG, Y.; YANG, Q. A survey on multi-task learning. **IEEE transactions on knowledge and data engineering**, IEEE, v. 34, n. 12, p. 5586–5609, 2021. Citado 2 vezes nas páginas 4 e 6.

ZHANG, Z.; VARADARAJAN, B. Utility scoring of product reviews. In: **Proceedings of the 15th ACM international conference on Information and knowledge management**. [S.l.: s.n.], 2006. p. 51–57. Citado na página 10.

AGRADECIMENTOS

Primeiramente, agradeço a Deus, por me conceder força, sabedoria e perseverança ao longo desta jornada acadêmica. Sem Sua graça e direção, este trabalho não teria sido possível.

Agradeço à minha mãe Maria da Conceição Vicente Pereira e ao meu pai Antonio Cláudio de Sousa, por todo o amor, apoio e sacrifício ao longo da minha trajetória acadêmica. Sem o exemplo de força e dedicação de vocês, essa conquista não seria possível. que sempre foram meu alicerce, pelo amor incondicional, apoio e incentivo em cada passo do meu caminho.

Agradeço também à Déborah Andrade, por estar sempre ao meu lado durante essa jornada estudantil, oferecendo apoio, incentivo e companheirismo nos momentos mais desafiadores.

Expresso minha gratidão especial a Ana Clara e Francisco Wesley, que se tornaram uma verdadeira família para mim quando eu mais precisei, especialmente nos momentos em que a distância da minha própria família foi mais sentida. O acolhimento e o carinho de vocês foram fundamentais para que eu seguisse em frente.

Ao meu orientador, Rogério Figueredo, expresso minha profunda gratidão pela paciência, orientação e por compartilhar seu conhecimento ao longo deste processo. Seu apoio e dedicação foram fundamentais para a realização deste trabalho. Por fim, agradeço a todos que, direta ou indiretamente, contribuíram para a concretização deste estudo. Seja através de palavras de incentivo, apoio técnico ou companhia nas horas difíceis, cada gesto fez a diferença nesta caminhada.

Muito obrigado!