Highlights

Explorando os Métodos Classificação de Politeness, utilizando o dataset PolitePEER

Augusto Giani, Fernando Farias

- Classificação de politeness utilizando métodos de machine learning clássicos
- Utilização de diferentes técnicas de pré-processamento
- Avaliação da performance dos modelos

Explorando os Métodos Classificação de Politeness, utilizando o dataset PolitePEER

Augusto Giani, Fernando Farias

Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil.

ABSTRACT

This paper investigates the automatic classification of tone in academic reviews using the PolitePEER dataset [1]. Preprocessing techniques, such as TF-IDF for text vectorization, were applied, and several supervised machine learning algorithms were utilized, including Logistic Regression, K Neighbors Classifier, and Naive Bayes. To address class imbalance and interpret the results, balancing techniques and explainability methods, such as SHAP, were implemented to identify the most relevant features in the predictions. The models were evaluated using metrics like F1-Score Macro and AUC-ROC, demonstrating that classical approaches can be effective in tone classification tasks while providing interpretable insights into the linguistic patterns associated with different levels of politeness.

RESUMO

Este trabalho investiga a classificação automática de tom em revisões acadêmicas utilizando o dataset PolitePEER [1]. Aplicou-se técnicas de pré-processamento, como TF-IDF para representação vetorial dos textos, e utilizamos diversos algoritmos de aprendizado de máquina supervisionado, incluindo Logistic Regression, K Neighbors Classifier, Naive Bayes. Para lidar com o desbalanceamento das classes e interpretar os resultados, foram implementadas técnicas de balanceamento e explicabilidade, como SHAP, para identificar os atributos mais relevantes nas predições. Os modelos foram avaliados por métricas como F1-Score Macro e AUC-ROC, demonstrando que abordagens clássicas podem ser eficientes na tarefa de classificação de tons, ao mesmo tempo que oferecem insights interpretáveis sobre os padrões linguísticos associados a diferentes níveis de polidez.

PALAVRAS-CHAVE

Text Classification, NLP, Machine Learning, Naive Bayes, Computational Social Science.

1. Introdução

O tom da comunicação é um aspecto fundamental da interação humana, especialmente em contextos profissionais e acadêmicos. No processo de revisão por pares, a polidez desempenha um papel crucial, impactando a recepção das críticas pelos autores e o ambiente colaborativo da ciência. Estudos recentes no campo de Computational Social Science (CSS) têm buscado automatizar a análise de polidez, fornecendo insights sobre as dinâmicas interpessoais em comunicações textuais. Este trabalho utiliza o dataset PolitePEER, uma base de dados pioneira que categoriza sentenças de revisões acadêmicas em cinco níveis de polidez. O objetivo central deste estudo é explorar e comparar diferentes abordagens para classificar automaticamente o tom dessas sentenças, com ênfase em algoritmos de aprendizado de máquina clássicos, suas comparações de performance e avaliação de explicabilidade.

Para alcançar esse objetivo, aplicou-se uma série de metodologias que vão desde a engenharia de atributos até a construção e explicação de modelos preditivos. Inicialmente, o dataset foi pré-processado com One-Hot Encoding (OHE) para variáveis categóricas e técnicas de balanceamento para lidar com o desbalanceamento entre as classes. Em seguida, utilizamos Term Frequency-Inverse Document Frequency (TF-IDF) para transformar os textos em representações vetoriais, priorizando

palavras-chave específicas de cada sentença e eliminando o peso de termos muito frequentes. Para a modelagem, treinamos diversos algoritmos clássicos, como Logistic Regression, Multinomial Naive Bayes e XGBoost, além de ajustar hiperparâmetros de um modelo de Random Forest. Adicionalmente, utilizamos a técnica SHAP para explicabilidade dos modelos, identificando as palavras que mais contribuíram para as predições.

Os resultados não apresentaram performances significativamente positivas, porém destacam-se Logistic Regression, XGBoost e Random Forest, que apresentaram desempenho melhor em métricas como F1-Score Macro e AUC-ROC, com destaque para a estabilidade do Multi-layer Perceptron (MLP) Classifier. O modelo de Random Forest ajustado também foi avaliado com explicabilidade via SHAP, permitindo uma análise mais aprofundada sobre as contribuições dos atributos.

2. Contexto

A polidez (*politeness*) é um aspecto essencial da comunicação humana, refletindo o esforço para manter interações respeitosas e socialmente adequadas. No campo de CSS, a polidez é estudada como um componente central das dinâmicas interpessoais, sendo usada para entender como a linguagem influencia relações, conflitos e colaboração em diferentes contextos e ambientes de interação. A análise automatizada de polidez permite identificar padrões linguísticos associados a tons positivos e negativos, fornecendo insights sobre o comportamento humano em interações textuais.

Aplicações em CSS frequentemente utilizam modelos de aprendizado de máquina para classificar tons de polidez e entender como variam em diferentes contextos culturais, profissionais e sociais. Especificamente em revisões acadêmicas, a polidez assume um papel crítico, influenciando não apenas a recepção das críticas pelos autores, mas também o ambiente ético do processo de revisão. Estudos como o **PolitePEER** [1] avançaram ao criar datasets anotados que capturam nuances de tons em revisões, permitindo análises detalhadas e automáticas dessa característica.

3. Dataset

O dataset PolitePEER é uma coleção pioneira de 2.500 sentenças extraídas de revisões acadêmicas em conferências renomadas, como ICLR e NeurIPS, bem como de fontes como o sistema PubIons e o blog ShitMyReviewersSay [1]. Este corpus foi desenvolvido com o objetivo de analisar e classificar o tom (Tone) das revisões em cinco categorias: Highly Impolite (altamente impolido), Impolite (impolido), Neutral (neutro), Polite (polido) e Highly Polite (altamente polido).

O corpus reflete a predominância de tons neutros (1.140 exemplos), seguidos de Impolite (582 exemplos) e Polite (465 exemplos), enquanto as classes extremas (Highly Impolite e Highly Polite) possuem respectivamente 210 e 103 exemplos, representando um desafio para classificação devido ao desbalanceamento.

As sentenças abrangem tópicos técnicos variados, características de revisões científicas em IA e áreas correlatas. Apesar disso, o dataset está limitado a um conjunto específico de disciplinas e não representa a diversidade cultural ou linguística global. As sentenças variam em comprimento, com uma média de 21 palavras, permitindo a análise de tons em diferentes contextos linguísticos e sintáticos.

O corpus foi anotado manualmente por quatro especialistas, seguindo diretrizes rigorosas que definem os níveis de polidez com base em marcadores linguísticos, atitudes expressas e convenções sociais. Discrepâncias nas anotações foram resolvidas por consenso em discussões entre os anotadores, resultando em uma alta concordância interavaliadora de 93%.

As definições das categorias foram detalhadas para capturar nuances, com classes extremas envolvendo julgamentos claros de impolidez (e.g., insultos ou críticas desnecessariamente ofensivas) ou polidez (e.g., elogios explícitos ou suavizações linguísticas).

O Tone refere-se ao estilo comunicativo adotado pelo revisor ao fornecer comentários. Ele varia de interações negativas e desrespeitosas (impolite) a comentários respeitosos e encorajadores (polite). A categorização busca capturar como a escolha de palavras, estrutura das frases e atitudes implícitas influenciam o impacto emocional e profissional da revisão. Na figura abaixo, é possível ver o desbalanceamento das classes do dataset:

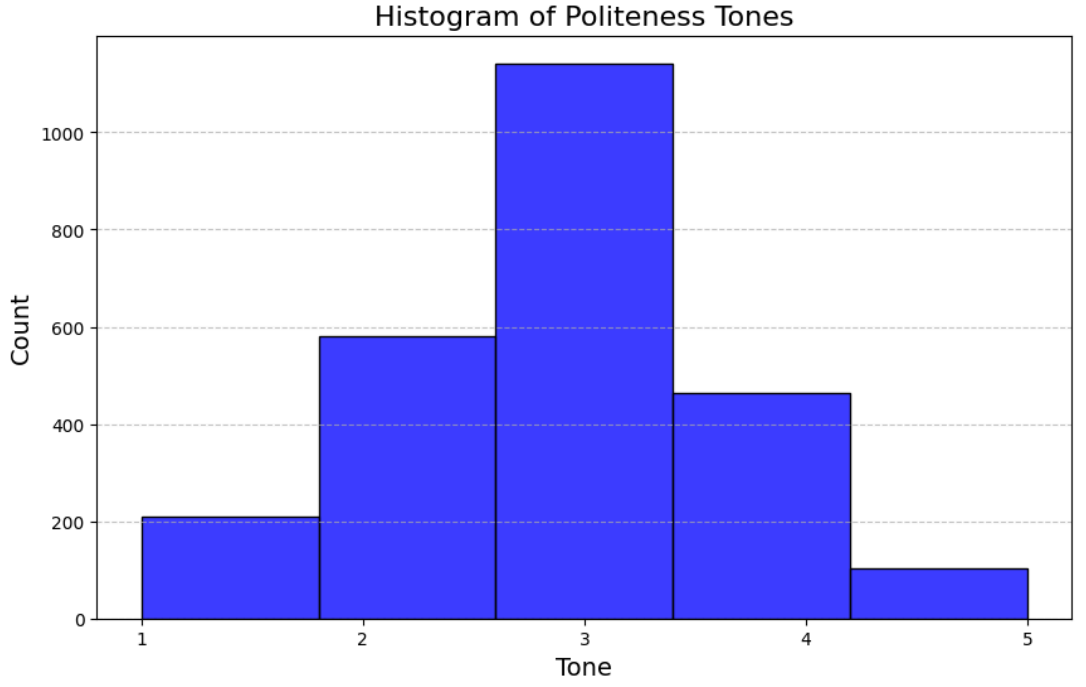

Figura 1. Distribuição dos Tones no dataset

4. Trabalhos Relacionados

O trabalho é baseado no artigo *PolitePEER: does peer review hurt? A dataset to gauge politeness intensity in the peer reviews, BHARTI, Prabhat Kumar et al.* [1], em que os autores abordam o problema da falta de polidez em revisões acadêmicas. Comentários impolidos e ásperos, muitas vezes desnecessariamente ofensivos ou condescendentes, podem impactar negativamente a confiança e o bem-estar dos autores, além de **comprometer a colaboração científica**. Para enfrentar esse problema, o estudo apresenta o *PolitePEER*, o dataset em questão, anotado em cinco níveis de polidez. Além disso, o artigo avalia diversos modelos de aprendizado de máquina, como SciBERT, ToxicBERT e um modelo customizado (Word2Vec + BiLSTM), para classificar automaticamente o tom das revisões. O objetivo é desenvolver ferramentas que alertem revisores sobre tons inadequados e incentivem feedback mais construtivo.

Apesar de suas contribuições, o artigo apresenta limitações significativas. O dataset, embora inovador, é pequeno e bem desbalanceado, o que prejudica a generalização dos modelos treinados. O estudo também não explora a avaliação de modelos clássicos de classificação, nem oferece análises qualitativas detalhadas sobre os erros dos modelos. Entre outros gaps, estes abrem oportunidades para

expandir o impacto do trabalho, seja por meio de estudos entre domínios, análise de interpretabilidade ou desenvolvimento de novos métodos utilizando o dataset.

Em outros estudos [2][3], é avaliado frequentemente o impacto do Processamento de Linguagem Natural (em inglês, NLP) na revisão de pares identificando e classificando polidez e revisando sua utilidade, uso de informações sigilosas. Também, em questão mais ampla, quais são os métodos que podemos aplicar para aprimorar a interação humana mais baseada em dados e inteligência artificial

5. Objetivo

- O objetivo deste trabalho é explorar diferentes algoritmos de aprendizado de máquina para a classificação automática de Tone em sentenças de revisões acadêmicas, utilizando o dataset PolitePEER.

- Adicionalmente, o trabalho visa interpretar as distribuições de palavras e padrões linguísticos nas decisões dos modelos por meio de técnicas de explicabilidade, promovendo uma análise mais transparente e robusta sobre o impacto de cada abordagem.

Não é objetivo deste trabalho avaliar as particularidades de cada modelo quanto às suas características intrínsecas, dado o dataset utilizado, apenas aplicá-los para avaliar suas performances com estes dados, tampouco as particularidades de cada método de avaliação, tendo em vista que são métodos já consolidados para os seus propósitos, aqui aplicados sem nenhuma exceção às suas aplicações gerais.

6. Metodologia

Neste trabalho, aplicou-se algumas técnicas de aprendizado de máquina para classificar automaticamente o Tone. O objetivo principal foi comparar o desempenho dos algoritmos clássicos de aprendizado de máquina:

- Logistic Regression;
- K-Neighbors Classifier;
- Bernoulli Naive Bayes,
- Multinomial Naive Bayes,
- Multi-layer Perceptron Classifier,
- Random Forest Classifier,
- XGBoost Classifier,

A aplicação destes modelos se deu para avaliar a eficácia na classificação de Tone. Além da comparação entre os modelos, também focou-se na explicabilidade dos modelos e na interpretação dos resultados utilizando SHAP, para entender quais características mais contribuíram para as previsões. As metodologias aplicadas envolveram desde o pré-processamento e engenharia de características até o ajuste de modelos e a avaliação, garantindo uma análise mais completa do problema, se principalmente comparado ao baseline de quanto o dataset foi proposto.

Foi utilizada a estratégia de One-Hot Encoding (OHE) para o treinamento dos modelos e avaliação das performances.

A sequência temporal de experimentos foi feita na ordem respectiva do pré-processamento, depois foi-se treinado os modelos, na sequência foram-se ajustados os hiperparâmetros dos modelos selecionados, depois avaliados os modelos ajustados e por mim realizado a explicabilidade dos modelos ajustados. Todos os métodos estão descritos abaixo.

6.1. Pré Processamento

A primeira etapa deste trabalho foi o pré-processamento, no qual foram aplicadas as seguintes técnicas:

1. Remoção de caracteres especiais e lemmatização
2. OHE para a variável a ser predita (tone)
    a. O One-Hot Encoding é uma técnica de pré-processamento usada para transformar variáveis categóricas em uma representação numérica binária. Foi criada uma coluna para cada tone (*tone_1, tone_2, tone_3, tone_4, tone_5*).
3. Train and Test Split
    a. Foi se dividido os dados na proporção 67% treino e 33% teste
    b. Aqui foi usado método *stratify* para balanceamento de classes, em que a divisão é feita baseando-se na proporção das classes.
4. Método de balanceamento de classes:
    a. Foi utilizado o SMOTE nos dados de treinamento para lidar com desbalanceamento e ajudar o modelo a aprender melhor os padrões da classe minoritária, realizando upsampling. O conjunto de testes permaneceu intacto, sem qualquer alteração.
5. TF-IDF para:
    a. Garantir que as palavras mais comuns não apareçam com palavras-chaves no treinamento dos modelos
    b. Mantém palavras frequentes específicas do documento com peso maior.
    c. Os vetores TF-IDF foram construídos para o treinamento dos modelos, filtrando com um mecanismo RegEx que busca palavras compostas exclusivamente por letras (sem números ou símbolos) e que tenham pelo menos dois caracteres.

6.2. Treinamento

O treinamento dos modelos deu-se iterativamente para cada Tone, no intuito de avaliar-se a performance de cada modelo para cada Tone utilizando a estratégia de OHE,

Os seguintes modelos treinados usaram os seguintes hiperparâmetros:

```
LogisticRegression, {'random_state': 42},
KNeighborsClassifier, {'n_neighbors': 5},
BernoulliNB, {},  # No parameters for BernoulliNB
MultinomialNB, {}, # No parameters for MultinomialNB
MLPClassifier,    {'hidden_layer_sizes':    50,    'solver':    'lbfgs',    'max_iter':    10000,
'random_state': 42},
RandomForestClassifier,  {'n_estimators':  50,  'max_depth':  None,  'min_samples_split':  2,
'random_state': 42},
XGBClassifier, {} # No parameters for XGBClassifier
```

6.3. Ajustes do Random Forest Classifier e Multinomial NB

O modelo Random Forest Classifier, devido ao seu desempenho destacado, foi explorado mais além para ajuste de seus hiperparâmetros e para melhor avaliação dos métodos aplicados.

Seus ajustes foram feitos através de um método de busca de melhores hiperparâmetros usando Cross-Entropy Loss, para assim encontrar os seguintes hiperparâmetros para os modelos:

```
RandomForestClassifier(n_estimators=368,
                       max_depth=22,
                       min_samples_split=3,
                       min_samples_leaf=1,
                       ccp_alpha=0,
                       random_state= 42)
```

Também foi-se usado Grid Search para melhorar os hiperparâmetros do modelo **MultinomialNB**, foi-se utilizado os seguintes parâmetros de procura:

- `alpha = [0.1, 0.5, 1.0, 2.0, 5.0]`
  E o ajuste de modelos após otimização de hiperparâmetros foi com `alpha = 0.1`.

6.4. Avaliação

Para cada modelo Random Forest Classifier e Multinomial Naive Bayes treinado foram-se avaliados a curva ROC, ROC-AUC Score, Matriz de Confusão e F1-Macro, de forma a ter parâmetros conhecidos e já consolidados para avaliar o quanto um modelo ajustado com certos hiperparâmetros performou melhor do que outro.

6.5. Explicabilidade

A avaliação de explicabilidade foi feita através do método SHAP (SHapley Additive exPlanations), que visa interpretar modelos de aprendizado de máquina, atribuindo a contribuição de cada característica às previsões do modelo (em linhas gerais, o método utiliza conceitos da teoria dos jogos cooperativos para calcular quanto cada característica contribui, de forma positiva ou negativa, para a previsão). Também, foram-se construídas as Nuvens de Palavras, das 100 palavras que mais contribuíram para cada modelo Multinomial Naive Bayes e Random Forest Classifier ajustado.

7. Resultados

7.1. Treinamento dos modelos

No primeiro treinamento dos modelos, foi-se avaliado suas performances em relação à ROC AUC, como é possível observar na Figura 2. Na implementação, também foram-se avaliados a acurácia e f1-score dos modelos, evidenciando resultados na média de 0,5 - 0,6 para maioria dos modelos.

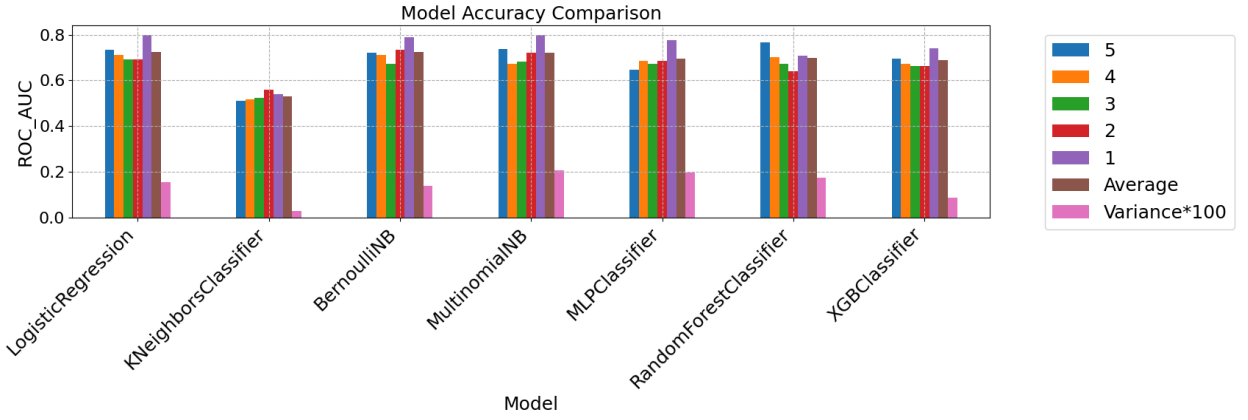

Figura 2. Resultados de cada modelo em relação ao ROC AUC Score, ajustado para cada Tone utilizando OHE. Os números de 1 - 5 representam os modelos ajustados para cada Tone, e.g. MLPClassifier 1 significa o ROC AUC Score do modelo ajustado para o Tone 1 pela estratégia OHE.

Como é possível observar, já no treinamento os modelos apresentaram performance de acurácia (refletida no ROC AUC Score) baixa, com a maioria na faixa entre 0,4 - 0,6 e 0,6 - 0,8.

Tone Level	**5**	**4**	**3**	**2**	**1**	**Average**	**Variance*100**
Model							
LogisticRegression	0,7336	0,7097	0,6910	0,6930	0,7978	0,7250	0,1298
KNeighborsClassifier	0,5115	0,5178	0,5245	0,5585	0,5409	0,5307	0,0242
BernoulliNB	0,7196	0,7101	0,6729	0,7339	0,7873	0,7248	0,1154
MultinomialNB	0,7363	0,6712	0,6826	0,7224	0,7994	0,7224	0,1721
MLPClassifier	0,6460	0,6851	0,6738	0,6854	0,7773	0,6935	0,1633
RandomForestClassifier	0,7649	0,7030	0,6719	0,6394	0,7078	0,6974	0,1452
XGBClassifier	0,6950	0,6722	0,6621	0,6641	0,7419	0,6870	0,0740

Tabela 1. Relação de ROC-AUC Score para cada modelo ajustado para cada Tone utilizando OHE.

7.2. Avaliações dos modelos com hiperparâmetros ajustados

Random Forest Classifier

Random Forest Classifier Tone 1 obteve *ROC-AUC Score = 0,736* e *F1-Macro = 0,028*.

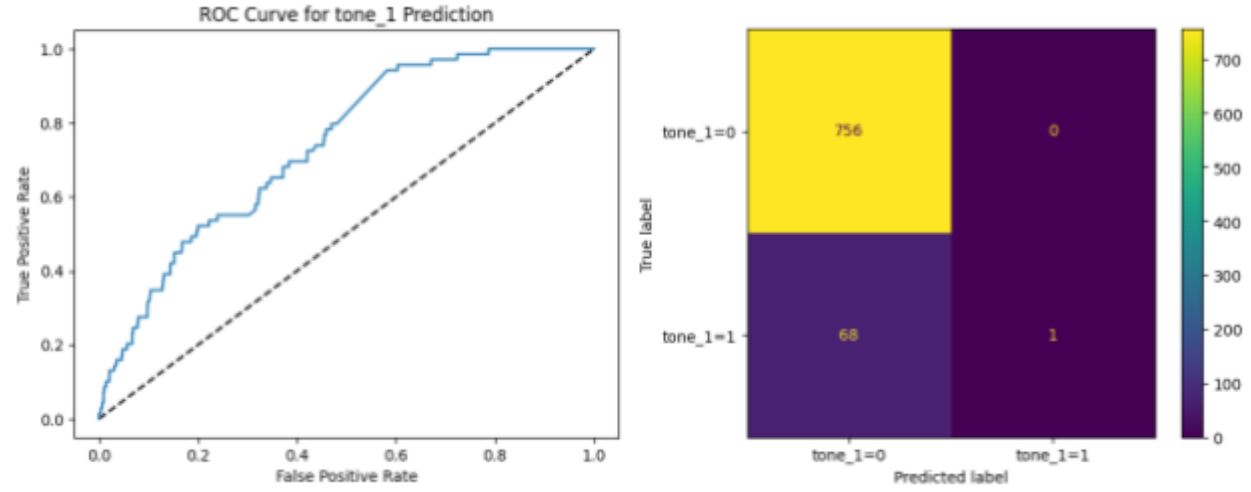

Figura 3. Curva ROC e Matriz de Confusão para Random Forest Classifier Tone 1.

Para este modelo ajudado para este dataset, é possível observar, na Figura 3, que ele erra muito por falsos negativos. Tone 1 é a classe Highly Impolite, que trás reviews com palavras mais ásperas e mais características, o que pode ter ajudado o modelo a se ajustar melhor para estes dados.

Random Forest Classifier Tone 2 obteve *ROC-AUC Score = 0,681* e *F1-Macro = 0,203.*

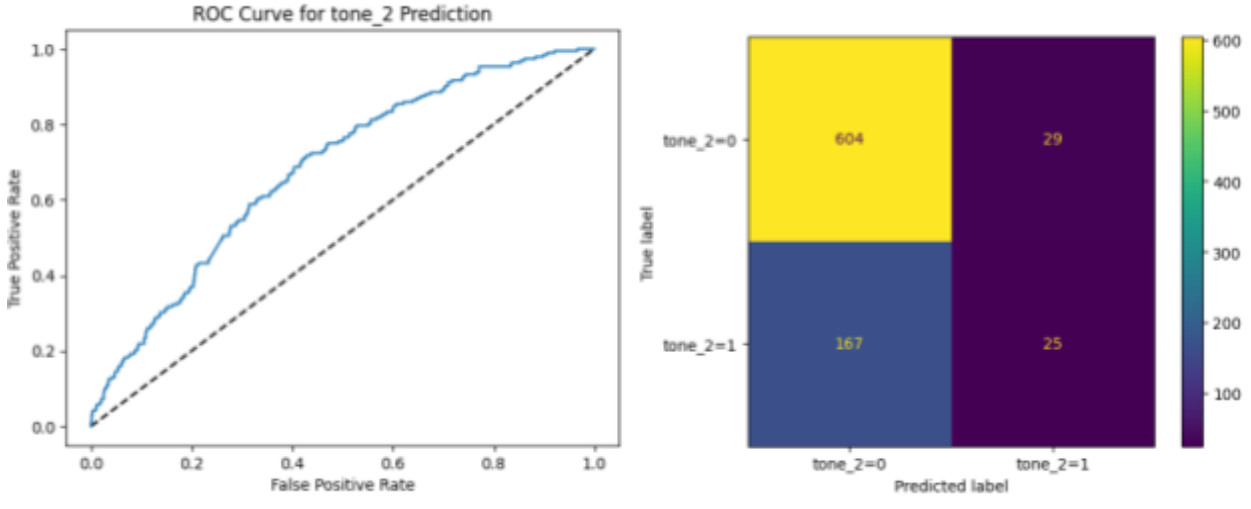

Figura 3. Curva ROC e Matriz de Confusão para Random Forest Classifier Tone 2.

O Tone 2 se trata de comentários Impolite, o que pode trazer comentários bem sutis sobre a review em questão, o que pode dificultar a performance do modelo, com é possível ver a quantidade de predições erradas na Figura 3.

Random Forest Classifier Tone 3 obteve *ROC-AUC Score = 0,688* e *F1-Macro = 0,579*

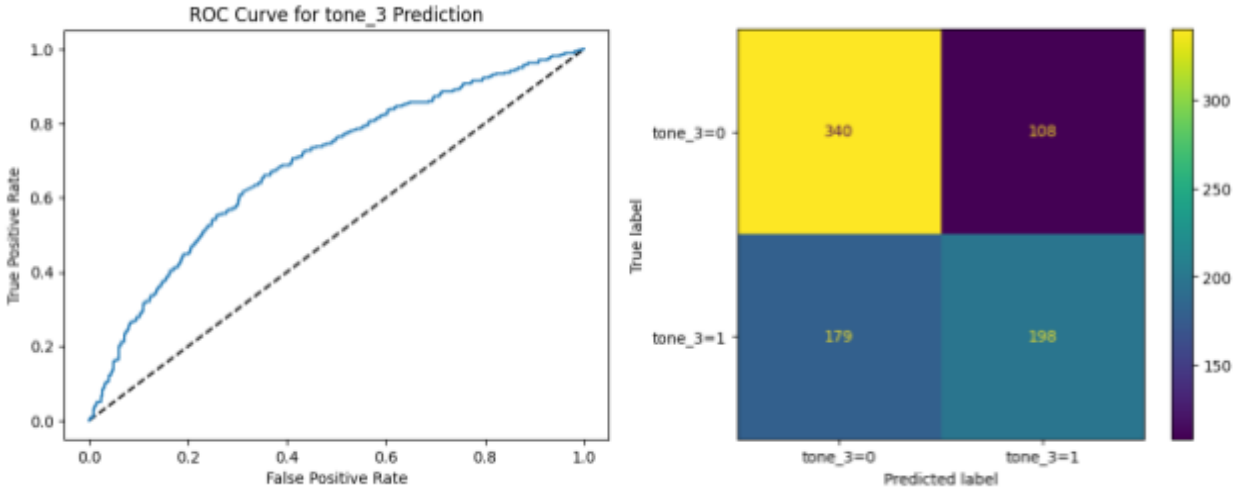

Figura 4. Curva ROC e Matriz de Confusão para Random Forest Classifier Tone 3.

O Tone 3 traz os comentários mais comuns e com a menor variedade de vocabulário, por isso é esperado a maior quantidade de predições erradas para este modelo. Importante ressaltar a interpretação sobre o F1-Score, que leva em consideração a acurácia do modelo e que não reflete muito bem a real performance do modelo ao se avaliar esta métrica sozinha.

Random Forest Classifier Tone 4 obteve *ROC-AUC Score = 0,722* e *F1-Macro = 0,125*

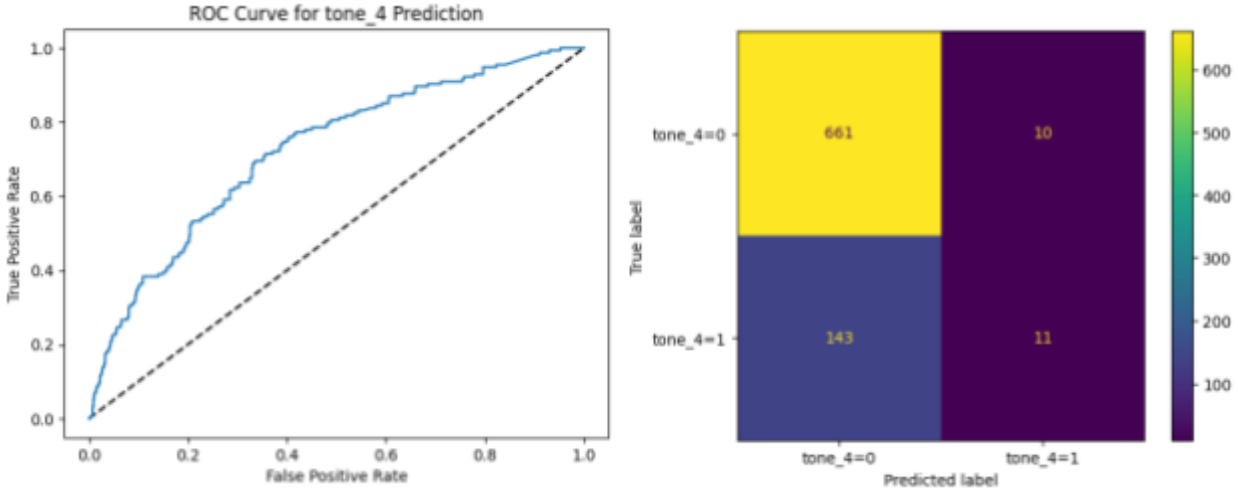

Figura 5. Curva ROC e Matriz de Confusão para Random Forest Classifier Tone 4.

É possível ver similaridades da performance do modelo do Tone 4 com do Tone 2 devido à estar distante na mesma proporção, são comentários Polite e também o dataset possuia praticamente a mesma proporção de dados desta classe.

Random Forest Classifier Tone 5 obteve *ROC-AUC Score = 0,707* e *F1-Macro = 0*

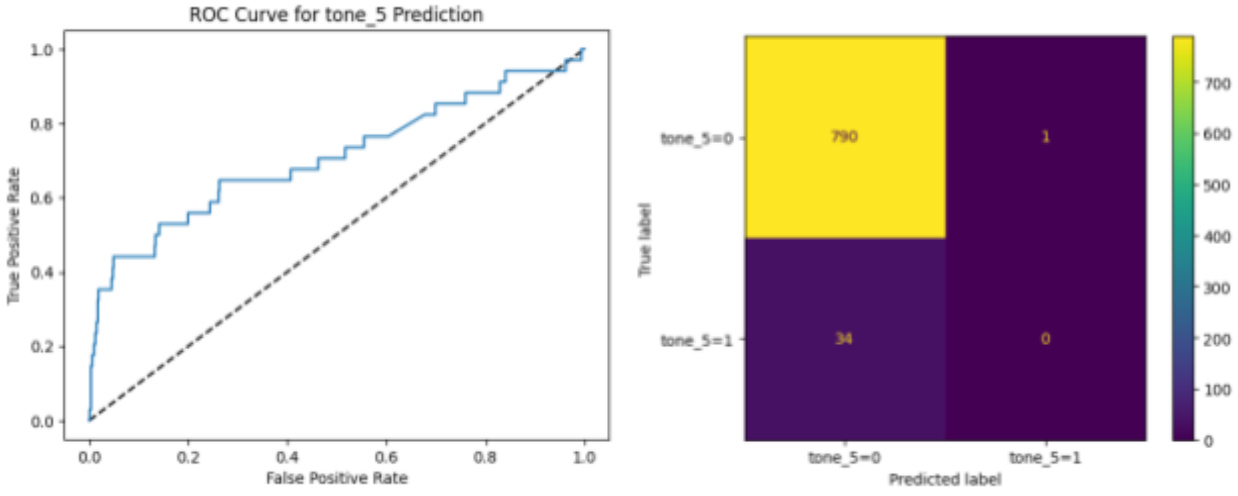

Figura 6. Curva ROC e Matriz de Confusão para Random Forest Classifier Tone 5.

O modelo utilizando OHE para o Tone 5 (Highly Polite - HP) praticamente colocava todos os exemplos para falsos e acertou para aqueles, porém não conseguia predizer nenhum para HP.

7.3. Multinomial Naive Bayes

Também foi-se ajustado o modelo Multinomial Naive Bayes e avaliado da mesma forma que os modelos de Random Forest, Obtiveram-se os seguintes resultados:

Model	F1-Macro	ROC-AUC Score
MultinomialNB Tone 1	0,8019	0,3121
MultinomialNB Tone 2	0,7261	0,4474
MultinomialNB Tone 3	0,6607	0,5976
MultinomialNB Tone 4	0,6136	0,3227
MultinomialNB Tone 5	0,6933	0,1684

Tabela 2. Resultados do ROC AUC Score e F1-Macro para modelos treinados baseados em Naive Bayes.

Os resultados dos modelos Multinomiais Naive Bayes foram ligeiramente melhores em relação aos modelos Random Forest com hiperparâmetros ajustados.

7.4. Nuvens de Palavras

Para melhor avaliação das palavras que mais contribuíram para predição, foram-se construídas as nuvens de palavras de todos os modelos com hiperparâmetros ajustados. Destacam-se algumas nuvens de palavras, como a das palavras mais relevantes para o modelo de MultinomialNB Tone 1, observando a figura abaixo.

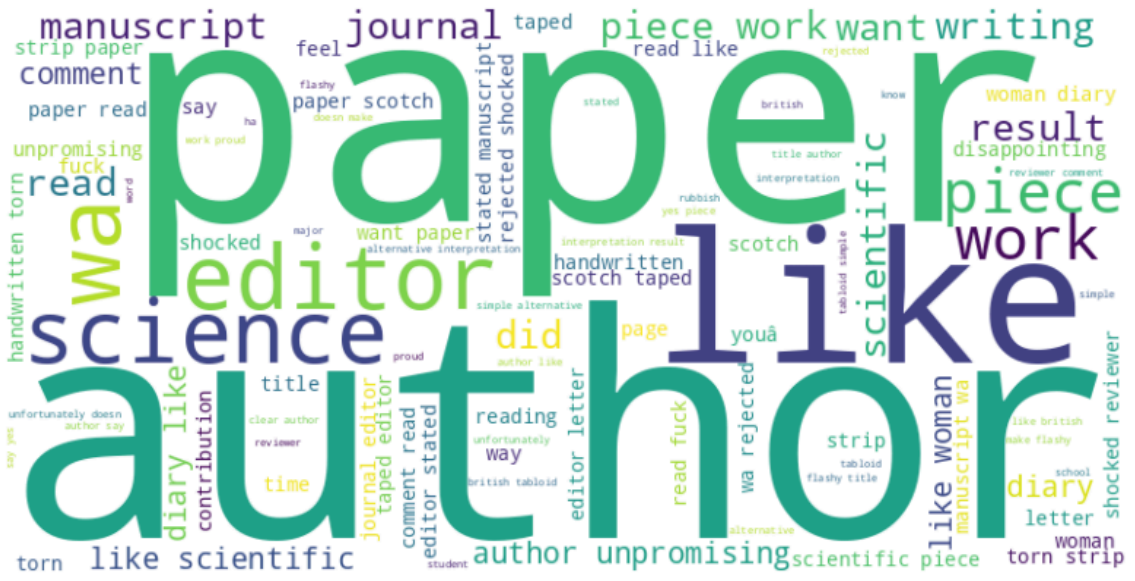

Figura 7. Nuvem de palavras das 100 palavras mais relevantes na classificação em um modelo Multinomial Naive Bayes Tone 1.

Também, para o modelo Multinomial Naive Bayes Tone 5, na Figura 8.

Figura 8. Nuvem de palavras das 100 palavras mais relevantes na classificação em um modelo Multinomial Naive Bayes Tone 5.

É possível observar que para a nuvem de palavras do Tone 1 se encontram alguns termos que podem ser frequentes em comentários Impolite, e.g. “author unpromising”, ”torn strip”, “rejected”. Da mesma forma, palavras que se encaixam mais em comentários mais polidos, como “thanks”, “dear”, “best”.

7.5. Explicabilidade Local

Os resultados em explicabilidade evidenciam as palavras que mais contribuem para a decisão de um modelo ao predizer uma nova instância. Destacam-se novamente os Tone 1 e 5, pelas suas diferenças marcantes nos estilos de comentários.

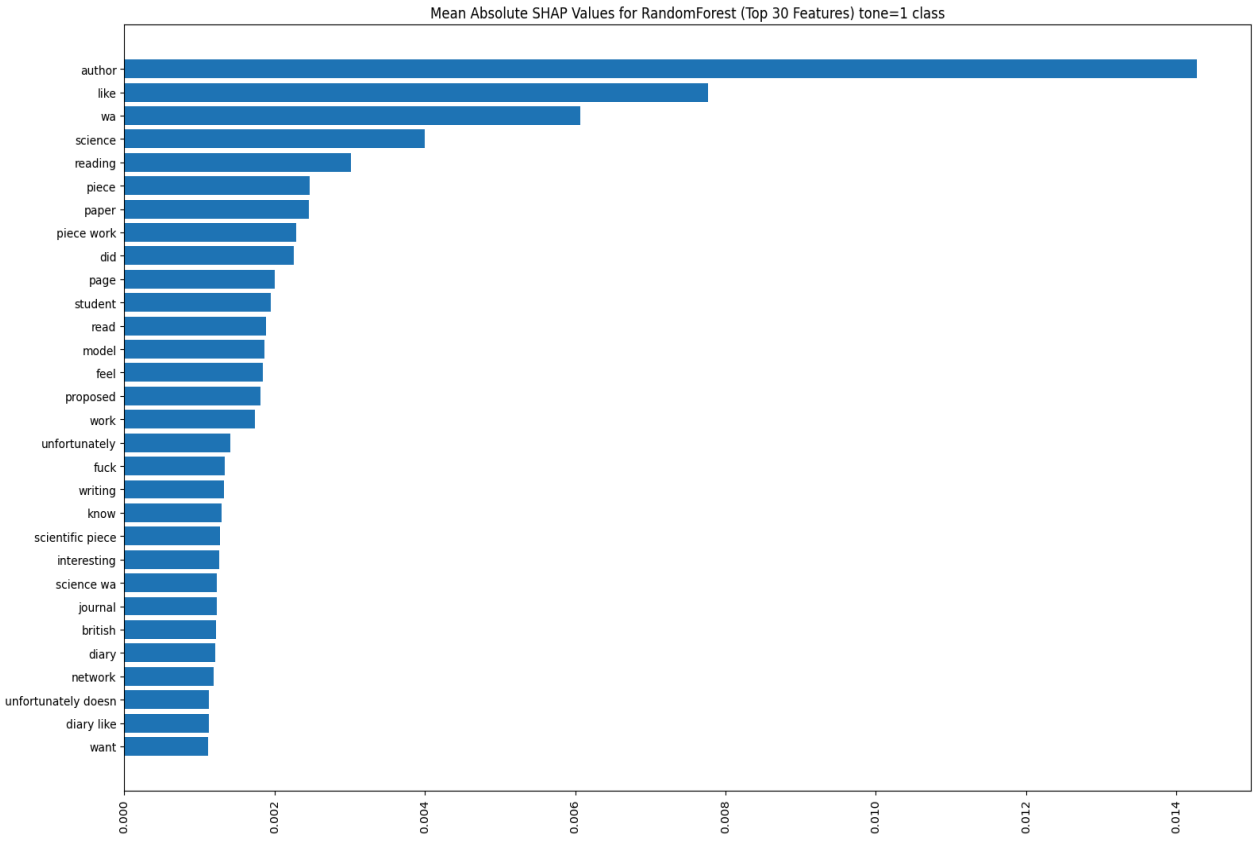

Figura 9. Gráfico SHAP para o modelo Random Forest Classifier Tone 1.

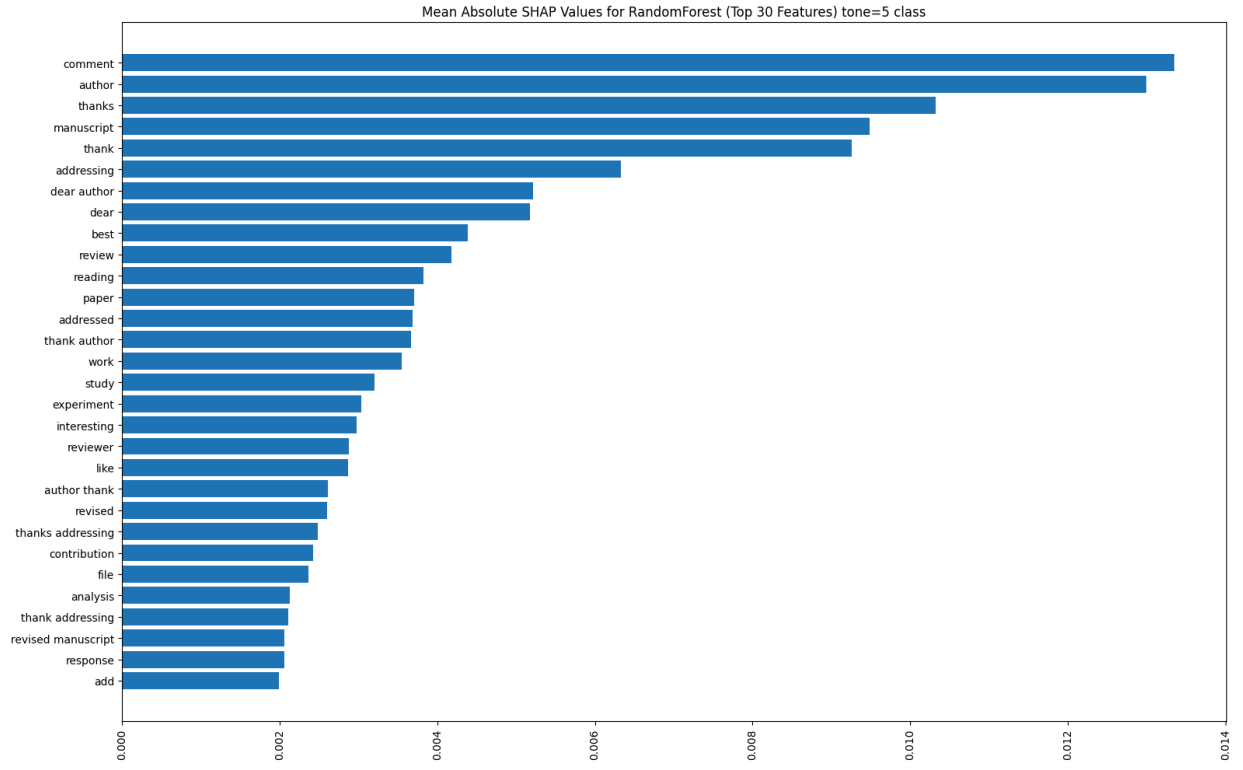

Figura 10. Gráfico SHAP para o modelo Random Forest Classifier Tone 5.

Para representar uma predição, foi utilizado o seguinte exemplo dos dados de teste: *"presentation clarity to the best of my understanding the author are some misconception"*. Probabilidade prevista de pertencer ao Tone 1: 13,12%, Classe prevista: False. Classe verdadeira: True.

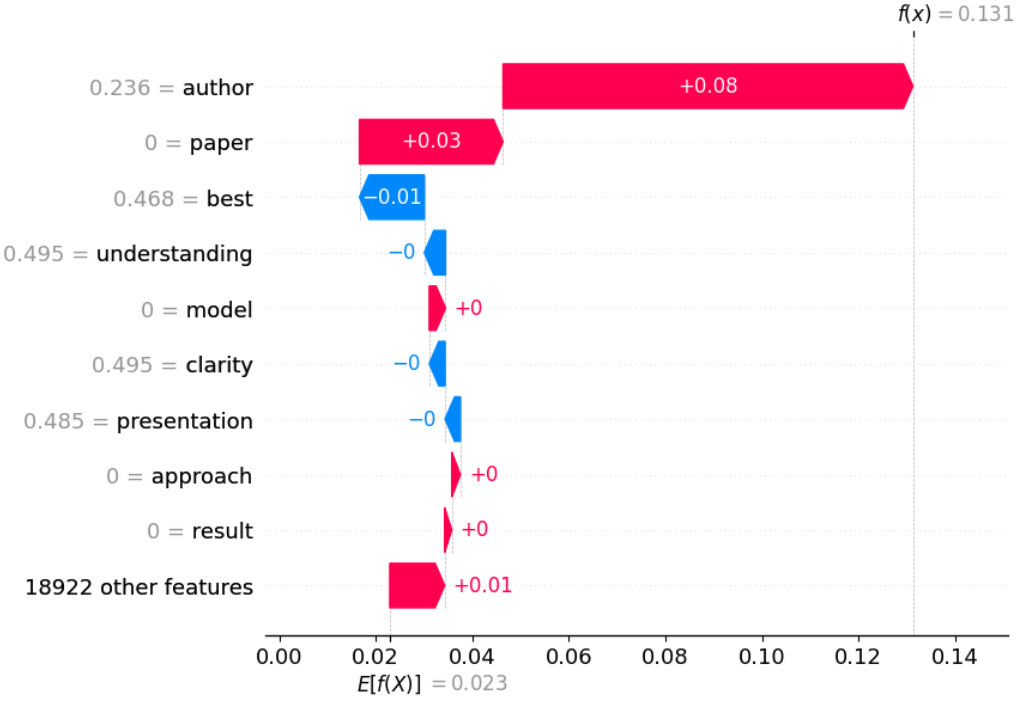

Figura 11. Explicabilidade da contribuição das features para a predição de uma review.

Existe um limiar do modelo que é setado para que seja considerado True (da classe Tone 1) se o valor da função f(x) for maior que 0,5, e isso explica o porque esta instância foi predita como False - não pertence à classe 1. Foi feito outros experimentos com outras reviews e é possível explicar melhor as predições utilizando esta estratégia

8. Conclusões

Este trabalho abordou o problema da predição de tom de polidez em textos, enfrentando desafios importantes como o desbalanceamento de classes e a escolha de representações apropriadas para os dados. Os resultados obtidos mostraram que:

- O desbalanceamento de classes teve um impacto significativo no desempenho dos modelos, mesmo com a utilização do método SMOTE para balanceamento dos dados. Isso evidencia a dificuldade de modelos em capturar nuances em classes menos representadas, destacando a necessidade de estratégias adicionais para melhorar a generalização.
- A inclusão de bi-gramas como representação textual foi um avanço relevante para capturar a semântica do problema de polidez, indo além das representações baseadas apenas em palavras isoladas (uni-gramas). Essa abordagem permitiu que os modelos tivessem uma visão mais contextualizada das relações entre palavras, resultando em uma melhor compreensão do tom dos textos.
- Embora tenham sido testadas várias arquiteturas de modelos, a performance geral refletiu as limitações impostas pela dificuldade de reprodução da baseline e o treinamento com um conjunto pequeno de reviews desde dataset, evidenciando a falta de informações como metadados e informações complementares ao dataset, podendo eventualmente serem acrescentados às features dos modelos. Apesar disso, as diferenças entre as abordagens analisadas sugerem potenciais caminhos de melhoria.

Com base nos desafios e resultados observados, algumas direções futuras podem ser consideradas para expandir e aprimorar este trabalho:

- Implementar um sistema de votação entre os cinco modelos treinados pode melhorar a robustez das predições, combinando as forças individuais de cada abordagem para mitigar suas limitações. Essa estratégia pode ajudar a equilibrar melhor as decisões em casos ambíguos ou complexos.
- Reformular o problema como uma tarefa de regressão pode ser uma alternativa interessante. Isso permitiria tratar a polidez como uma escala contínua, capturando melhor a gradação dos tons de polidez, ao invés de forçar as predições a classes discretas. Modelos regressivos poderiam explorar mais nuances nos dados, possivelmente aumentando a precisão e utilidade das predições.

Este trabalho contribuiu para o estudo de classificação de tom de polidez em textos, abordando questões técnicas e metodológicas que são relevantes para problemas de NLP em geral. A introdução de bi-gramas e a avaliação do impacto do desbalanceamento foram avanços importantes. As estratégias futuras delineadas oferecem um horizonte promissor para superar as limitações enfrentadas e expandir o potencial de aplicação dos resultados obtidos.

9. Referências

[1] BHARTI, Prabhat Kumar et al. **PolitePEER: does peer review hurt? A dataset to gauge politeness intensity in the peer reviews**. Language Resources and Evaluation, p. 1-23, 2023.

[2] RASTOGI, Charvi. **Improving Human Integration across the Machine Learning Pipeline**. 2024. Tese de Doutorado. Northwestern University.

[3] KUZNETSOV, Ilia et al. What Can Natural Language Processing Do for Peer Review?. **arXiv preprint arXiv:2405.06563**, 2024.