# Aprimorando a Eficiência e a Equidade de uma Abordagem Perspectivista Para Detecção de Ironia

Samuel B. Jesus
Federal University of Minas Gerais
Belo Horizonte, Brazil
samuelbrisio@dcc.ufmg.br

Guilherme D. Bianco
Universidade Federal da Fronteira Sul
Chapecó, Brazil
guilherme.dalbianco@uffs.edu.br

Wanderlei Junior
Federal University of Minas Gerais
Belo Horizonte, Brazil
wanderlei-junior@ufmg.br

Valerio Basile
University of Turin
Turin, Italy
valerio.basile@unito.it

Marcos André Gonçalves
Federal University of Minas Gerais
Belo Horizonte, Brazil
mgoncalv@dcc.ufmg.br

## ABSTRACT

Text classification in tasks like hate speech or irony detection is culturally influenced and personally interpretive. Perspectivism, unlike approaches that aggregate opinions by, for instance, majority voting, values specific annotator groups to create fairer models but often involves high computational costs due to fine-tuning of language models. This work explores traditional machine learning methods (SVM, Random Forest, XGBoost) to reduce costs and uses calibration to address inference biases. Results show up to 12 times faster processing without statistical effectiveness loss and improved fairness through reduced bias.

## KEYWORDS

perspectivismo, redes neurais, combinação, detecção de ironia, calibração

## 1 INTRODUÇÃO

Em tarefas subjetivas, como detecção de discurso de ódio ou ironia, a classificação de textos dependem do conhecimento cultural e do impacto individual do discurso em cada indivíduo [3]. Uma característica inerente desse tipo de problema é o desacordo de rótulo (*label disagreement*) (ódio vs. não-ódio ou irônico vs. não-irônico) [2]. Trata-se de um processo natural, decorrente de diferenças culturais e de como os indivíduos percebem ou são afetados por determinados discursos. Essa individualização da percepção, refletidas nos rótulos, pode fornecer informações valiosas para a tarefa de detecção (classificação) automática de discurso.

Métodos tradicionais de classificação agregam múltiplas anotações por meio de estratégias como a escolha da classe majoritária, descartando visões minoritárias ou menos representativas [5]. A proposta do **perspectivismo** é preservar as múltiplas anotações para capturar diferentes visões, promovendo maior equidade [6]. Ao treinar modelos independentes por grupo cultural, cada um refletindo interpretações específicas, considera-se a diversidade cultural nos dados. Fundamentalmente, o perspectivismo busca mitigar vieses contra grupos historicamente marginalizados, como LGBTQ+, populações negras, indígenas e minorias religiosas [1]. Particularmente, em [4], propõe-se um método perspectivista com combinação (*ensemble*) de modelos ajustados por perspectiva, cujos resultados indicam combinações promissoras. A despeito dos bons resultados de efetividade, o ajuste fino de múltiplos de linguagem impõe elevada demanda computacional.

Neste contexto, este trabalho possui dois objetivos centrais. O primeiro é aumentar a *eficiência* da abordagem perspectivista de [4] — doravante denominada método base — por meio da integração com modelos tradicionais de aprendizado de máquina — SVM, Regressão Logística e XGBoot —, buscando manter efetividade com menor custo computacional. O segundo é aprimorar a *equidade* entre modelos perspectivistas por meio de técnicas de calibração.

No método base, observou-se que algumas perspectivas apresentaram baixa representatividade (baixa confiança nas predições), o que limita ou inviabiliza sua contribuição para o rótulo final, comprometendo o princípio de equidade do perspectivismo. Hipotetizamos que tal efeito decorre da descalibração[1] dos modelos, o que pode resultar em probabilidades incompatíveis ou subestimadas. Assim, incorporou-se uma etapa de calibração para aumentar a confiabilidade dos métodos. Os resultados experimentais demostram que a combinação com modelos tradicionais reduz o tempo de execução em até 12 vezes, sem perda estatística na eficácia. A calibração também promove maior equilíbrio na contribuição das diferentes perspectivas no resultado final, gerando modelos mais *justos* sob a ótica do perspectivismo.

## 2 TRABALHOS RELACIONADOS

Em tarefas subjetivas — como detecção de discurso de ódio, ironia, sentimentos e linguagem abusiva — a anotação por múltiplos julgadores é frequentemente necessária [7]. Nessas situações, a divergência entre rótulos costuma ser tratada como ruído [5], adotando-se o voto da maioria e desconsiderando perspectivas de grupos minoritários potencialmente afetados [1]. O perspectivismo propõe incorporar e valorizar a diversidade de interpretações presentes nos dados, modelando as variações individuais a partir de características culturais e demográficas dos anotadores [3]. Tal abordagem

[1] Em um modelo de classificação adequadamente calibrado, a probabilidade a posteriori estimada pelo classificador apresenta alta correspondência com a frequência empírica de acertos. Especificamente, se o modelo atribui uma probabilidade de 80% a uma classe para um conjunto de instâncias, espera-se que, aproximadamente, 80% dessas predições estejam corretas.

In: V Concurso de Trabalhos de Iniciação Científica (CTIC 2025). Anais Estendidos do XXXI Simpósio Brasileiro de Sistemas Multimídia e Web (CTIC'2025). Rio de Janeiro/RJ, Brasil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2596-1683

tem ganhado destaque diante da crescente demanda por modelos justos, inclusivos e sensíveis a vieses [1, 3, 5].

Em [4], cada perspectiva é modelada individualmente dividindo-se o conjunto de treino em subconjuntos correspondentes a grupos específicos (e.g., anotadores masculinos e femininos). O ajuste fino de modelos de linguagem é realizado em cada subconjunto para extrair padrões particulares, e as predições são posteriormente agregadas por métodos baseados em confiança, gerando uma predição única. [5] propõe avaliar a pontuação atribuída por cada anotador pertencente ao grupo-alvo do discurso de ironia, utilizando dois módulos em paralelo: GPT-2 para identificar o grupo-alvo e RoBERTa para estimar a pontuação do anotador, ambos ajustados para a tarefa. Já [12] busca capturar padrões individuais concatenando textos rotulados pelo mesmo anotador ao texto a ser inferido, codificando assim as crenças do anotador junto ao input para o modelo de linguagem.

## 2.1 Abordagem Proposta

Nesta seção, demonstraremos como abordagens tradicionais de aprendizado de máquina podem ser combinadas para aprimorar a eficiência do perspectivismo. Adicionalmente, apresentaremos como a calibração pode ser incorporada ao método para melhorar a equidade das perspectivas.

A Figura 1-Esquerda ilustra a abordagem perspectivista (método base) proposta em [4], composta por quatro etapas. Inicialmente, os dados de treino são particionados em subconjuntos perspectivistas com base nas informações dos anotadores (*Parte 1*), como sexo (masculino ou feminino) ou nacionalidade. Em seguida (*Parte 2*), são geradas representações densas para cada subconjunto utilizando um modelo de linguagem pré-treinado, com o principal custo computacional concentrado no ajuste fino. Na *Parte 3*, todos os modelos realizam inferência sobre o mesmo conjunto de teste. Por fim (*Parte 4*), as predições são agregadas por métodos tais como: (i) Confiança Máxima (CM), que adota a predição com maior confiança; (ii) Soma das Confianças (SC), que soma os escores de confiança; e (iii) Voto Majoritário, que considera a classe mais indicada entre as perspectivas.

A Figura 1-(b) apresenta adaptações ao método base, com modificações na Etapa2.b e adição da Etapa2.c. Na Etapa 2.b, substituem-se os modelos de linguagem por algoritmos tradicionais de classificação (SVM, Regressão Logística, XGBoost). Como esses métodos exigem entradas numéricas, utiliza-se um modelo de linguagem apenas para tokenização, sem ajuste fino (*zero-shot*), aproveitando sua capacidade de extrair padrões textuais complexos. As representações densas geradas alimentam os modelos tradicionais, que são posteriormente aplicados ao conjunto de teste, também representado densamente via modelo de linguagem. As demais etapas seguem o método base.

A calibração por *Platt Scaling*, utilizada neste trabalho, opera a partir da adição de um modelo de regressão logística sobre os escores (ou probabilidades) produzidos pelo classificador [10]. Ou seja, um novo modelo é gerado a partir do conjunto de validação para calibrar os pesos do método base. A Equação $P(y = 1 \mid s) = \frac{1}{1+e^{(A \cdot s + B)}}$ apresenta a função sigmoide utilizada no *Platt Scaling*, onde os parâmetros "A" e "B" são aprendidos, utilizando um conjunto de validação (treino), durante o ajuste de um modelo de regressão logística para calibrar as probabilidades. A intuição da equação é possibilitar que as probabilidades de entrada (ou *logits*) produzidos pelos modelos perspectivistas sejam ajustadas para refletirem a correta distribuição dos valores.

A calibração, exigiu a adição de um passo extra (Passo 2.c), que promove ajustes em probabilidades geradas pelos modelos de predição. A intuição é que com a calibração todas as probabilidades geradas pelas perspectivas tenham escalas similares evitando que uma perspectiva (descalibrada) domine o processo de geração de rótulos. Na Figura 1-(b) a, pode-se observar que a calibração utiliza as probabilidades geradas pelas inferências sobre o conjunto de validação (seta verde). A Parte 2.c é ortogonal ao modelo utilizado, ou seja, pode ser aplicado ao modelo de linguagem ou com os classificadores tradicionais.

# 3 AVALIAÇÃO EXPERIMENTAL

Nesta seção, apresentaremos os resultados obtidos a partir dos experimentos envolvendo os dois objetivos da pesquisa: (1) demonstrar o ganho de eficiência computacional com o uso de modelos tradicionais de ML no lugar do RoBERTa; e (2) e apresentar os impactos da calibração na equidade do método perspectivista. Os experimentos foram executados em um AMD 2990WX (64 threads, 3GHz), GeForce RTX 2080(8GB) e 128 GB de mémora RAM. O código fonte poderá ser acessado através do repositório do GitHub[2].

## 3.1 Conjunto de dados

Para a avaliação experimental, foi utilizado o *English Perspectivist Irony Corpus* (EPIC) [8]. O EPIC contém 3.000 registros de mensagens curtas oriundas do *Reddit* e do *Twitter*, rotuladas como irônicas ou não. Cada texto foi anotado, em média, por cinco indivíduos, permitindo a captura de variações associadas à geração, sexo e localização geográfica dos anotadores.

## 3.2 Métricas de avaliação

A eficácia é mensurada pelo *macro F1-score* [13], correspondente à média simples dos F1-scores por classe, atribuindo peso igual a todas. A eficiência é avaliada com base no tempo total (em segundos) equivalente á soma dos tempos dos processos de tokenização, treino, predição e calibração (quando aplicável), comparando-se abordagens com e sem perspectivismo. Cada experimento foi repetido 10 vezes com diferentes sementes, e os resultados incluem intervalo de confiança de 95% e análise estatística via teste de Wilcoxon com correção de Bonferroni para múltiplos métodos.

## 3.3 Resultados

A Tabela 1 apresenta a comparação entre o método base, utilizando o RoBERTa com ajuste fino, e métodos tradicionais de aprendizado de máquina, como Regressão Logística (RL), XGBoost e SVM. O RoBERTa e a Regressão Logística obtiveram os melhores valores de *F1-score*, com empate estatístico entre os métodos — exceto no Método de Confiança Máxima, no qual o RoBERTa apresentou um ganho estatístico de apenas 2.8 pontos percentuais. Em contraste, XGBoost e SVM tiveram desempenho inferior, com perdas superiores a 9%. O resultado superior da RL pode ser atribuído à sua capacidade de capturar relações lineares nos dados [11].

[2]https://github.com/dbguilherme/C-ENS_experiments/releases/tag/ACL_Article.

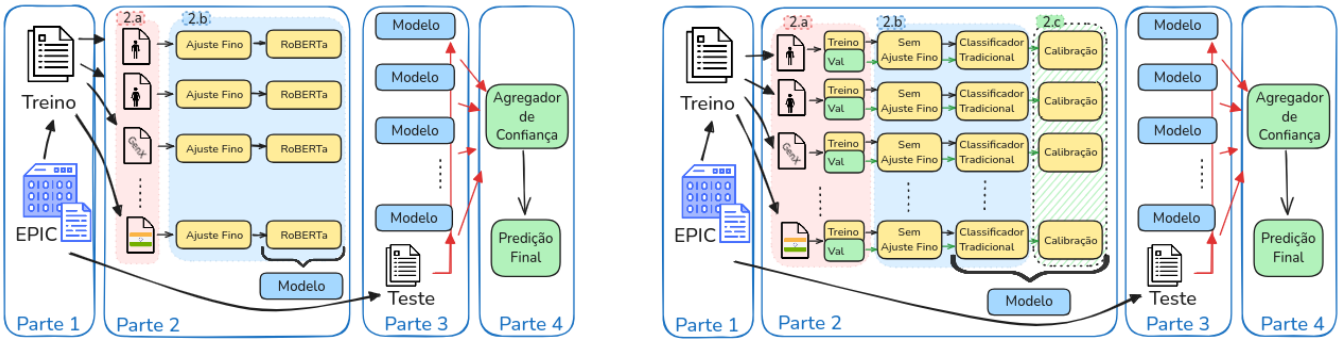


**Figura 1: Método perspectivista original [4] ilustrado à esquerda e o método com as alterações propostas à direita.**

| Sem Calibração | RoBERTa | Logística | XGB | SVM |
|---|---|---|---|---|
| Confiança Máxima (CM) | 67.4 ± 1.5 | 64.6 ± 1.2 ↓ | 53.6 ± 1.4 ↓ | 44.7 ± 0.6 ↓ |
| Somas das Confianças (SC) | 66.6 ± 1.7 | 65.2 ± 1.5 * | 54.1 ± 1.6 ↓ | 43.7 ± 0.4 ↓ |
| Voto Majoritário | 65.0 ± 2.0 | 64.3 ± 1.3 * | 54.0 ± 1.6 ↓ | 43.4 ± 0.2 ↓ |
| Sem-Perspectiva | 64.5 ± 2.5 | 63.3 ± 1.3 * | 57.1 ± 1.2 ↓ | 46.6 ± 0.9 ↓ |

**Tabela 1: F1-score (± IC) sem calibração para cada estratégia de agregação e modelo. '*' e '↓' representam empate ou perda estatística em relação ao RoBERTa.**

| Com Calibração | RoBERTa | Logística | XGB | SVM |
|---|---|---|---|---|
| Confiança Máxima (CM) | 67.0 ± 1.8 | 64.7 ± 1.1 * | 60.9 ± 1.3 ↓ | 63.1 ± 0.5 * |
| Somas das Confianças (SC) | 67.2 ± 1.7 | **65.2 ± 1.0** * | 62.9 ± 1.2 ↓ | 64.3 ± 0.9 * |
| Voto Majoritário | 67.0 ± 1.5 | 64.8 ± 1.5 * | 62.2 ± 1.3 ↓ | 64.4 ± 0.7 * |
| Sem-Perspectiva | 65.1 ± 1.7 | 62.1 ± 1.6 * | 60.4 ± 1.6 ↓ | 60.0 ± 1.4 ↓ |

**Tabela 2: F1-score (± IC) com calibração para cada estratégia de agregação e modelo. '*' e '↓' representam empate ou perda estatística em relação ao RoBERTa.**

A Tabela 2 apresenta os resultados com calibração, indicando melhorias em todos os modelos, exceto na Regressão Logística (RL), que já é calibrada por construção. Mas mesmo na RL, a calibração ajudou a reduzir a variância em alguns casos, especialmente à Soma da confianças. Os modelos tradicionais — *XGBoost* e *SVM* — foram os mais beneficiados, com aumentos de até 8.8 e 20.9 pontos percentuais no F1-score, respectivamente. No caso do RoBERTa com Voto Majoritário, observou-se uma melhora discreta, sem significância estatística. Destaca-se ainda que após a calibração, a RL obteve *empate estatístico com o RoBERTa em todos os casos.* Além disso, a calibração favoreceu o método de agregação por Soma das Confianças (SC), que obteve os melhores desempenhos em todos os classificadores.

| Tempo | RoBERTa | Logística | XGB | SVM |
|---|---|---|---|---|
| Sem-Perspectiva | 239.8 ± 13.7 | 16.5 ± 0.0 | 18.3 ± 0.1 | 22.8 ± 0.1 |
| Com-Perspectiva | 1904.7 ± 70.3 | 136.2 ± 0.5 | 154.1 ± 0.3 | 164.2 ± 0.5 |

**Tabela 3: Tempo de execução (em segs) com ICs de 95% sem a calibração.**

A Tabela 3 apresenta o tempo computacional das abordagens com e sem perspectivismo, utilizando RoBERTa e modelos tradicionais. Os métodos tradicionais são até 12 vezes mais rápidos que o RoBERTa. Comparando as Tabelas 3 e 4, observa-se que a calibração tem impacto mínimo no tempo de execução: aumento de apenas

| Time - Calibration | RoBERTa | Logística | XGB | SVM |
|---|---|---|---|---|
| Sem-Perpective | 245.2 ± 13.8 | 16.6 ± 0.1 | 18.9 ± 0.1 | 20.3 ± 0.1 |
| Perspectiva | 1951.4 ± 70.3 | 137.3 ± 0.4 | 161.0 ± 0.5 | 155.9 ± 0.8 |

**Tabela 4: Tempo de execução (em segs) com ICs de 95% utilizando a calibração.**

47 segundos (2%) para RoBERTa e 7 segundos (4%) para XGBoost. Já o SVM apresenta redução de 8.3 segundos (5%), possivelmente devido à diminuição do conjunto de treino, uma vez que parte dos dados é usada para calibrar via *Platt Calibration.*

A Figura 2 apresenta mapas de calor com a confiança das predições por perspectiva (554 linhas cada), sem calibração (à esquerda) e com calibração (à direita). Tons amarelos indicam baixa confiança; tons azuis, alta.

Os resultados revelam que quando não utilizamos calibração certas perspectivas (especificamente, *Boomer*, *Female*, e *GenY*) não têm influência significativa na previsão final, indicando que são efetivamente ignoradas pelo classificador e apenas introduzem custos computacionais desnecessários. Em contraste, o processo de decisão é dominado quase exclusivamente por duas perspectivas (*Ireland* and *GenX*) — Com o método de agregação CM, "Ireland" gera 48,9% das predições —, sugerindo um viés implícito em relação a essas dimensões. Compreender as razões para essa utilização desproporcional de perspectivas específicas constitui uma questão aberta intrigante, que deixamos como diretriz para investigações futuras.

Após a calibração, observa-se menor presença de tons amarelos, indicando maior participação de todas as perspectivas. "Ireland" reduz sua contribuição para 26,9%, enquanto "GenY" aumenta para 5%. Embora a calibração não traga ganhos estatísticos em todos os cenários, especialmente com modelo de linguagem RoBERTa, ela melhora a distribuição das predições, promovendo maior equidade entre as perspectivas, o que é consistente com o objetivo original do perspectivismo.

Resumindo, os experimentos ilustraram a possibilidade de redução do tempo de processamento de forma substancial sem perda estatística na eficácia, enquanto a calibração contribuiu para a geração de inferências mais justas, capazes de representar, de fato, os grupos minoritários, o objetivo principal do perspectivismo.

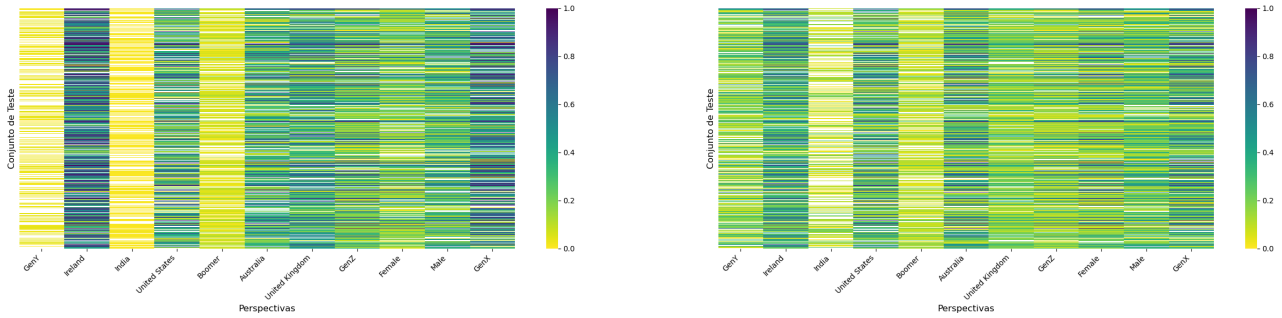


**Figura 2: Mapa de calor com a confiança de cada perspectiva sem (esquerda) e com a calibração (direita). O tom azul representa alta confiança do modelo na perspectiva; amarelo, baixa confiança. Modelos com alta confiança (tom azul escuro) dominam a geração de inferências.**

## 4 CONCLUSÃO

Propomos a integração de métodos tradicionais de classificação para aprimorar a eficiência de um método perspectivista recente. Verificou-se que, devido ao descalibramento das probabilidades, a abordagem de [4] gera predições enviesadas, sub-representando algumas perspectivas — em desacordo com seu objetivo central. Para mitigar esse efeito, incorporou-se a calibração como camada ortogonal, promovendo maior equidade e paridade em efetividade com o estado-da-arte.

Este trabalho possibilita que aplicações, como a moderação em redes sociais, sejam realizadas de forma justa, considerando a contribuição de cada perspectiva de maneira equitativa. Dessa forma, garante-se que grupos minoritários tenham participação semelhante à de grupos majoritários.

Como trabalho futuro, planeja-se explorar outros conjuntos de dados perspectivistas com o objetivo de verificar a abordagem em outros domínios, reforçando a generalização. E investigar técnicas de empilhamento [9] a fim de agregar os pontos fortes dos modelos tradicionais em uma única abordagem, propondo assim uma proposta mais robusta que possa superar os resultados da abordagem com LLM, visando maior efetividade com menor custo.

## REFERÊNCIAS


[1] Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose Opinions Matter? Perspective-aware Models to Identify Opinions of Hate Speech Victims in Abusive Language Detection. *CoRR* abs/2106.15896 (2021). arXiv:2106.15896 https://arxiv.org/abs/2106.15896

[2] L.M. Aroyo and C.A. Welty. 2015. Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *ACM Web Science 2013*.

[3] Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We Need to Consider Disagreement in Evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, Kenneth Church, Mark Liberman, and Valia Kordoni (Eds.). Association for Computational Linguistics, Online, 15–21. doi:10.18653/v1/2021.bppf-1.3

[4] Silvia Casola, Soda Marem Lo, Valerio Basile, Simona Frenda, Alessandra Teresa Cignarella, Viviana Patti, and Cristina Bosco. 2023. Confidence-based Ensembling of Perspective-aware Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 3496–3507. doi:10.18653/v1/2023.emnlp-main.212

[5] Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the Majority is Wrong: Modeling Annotator Disagreement for Subjective Tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6715–6726. doi:10.18653/v1/2023.emnlp-main.415

[6] Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation* (2024), 1–28.

[7] Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: A survey. *Language Resources and Evaluation* (2024). https://www.amazon.science/publications/perspectivist-approaches-to-natural-language-processing-a-survey

[8] Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: Multi-Perspective Annotation of a Corpus of Irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 13844–13857. doi:10.18653/v1/2023.acl-long.774

[9] Luca Gioacchini, Welton Santos, Barbara Lopes, Idilio Drago, Marco Mellia, Jussara M. Almeida, and Marcos André Gonçalves. 2024. Explainable Stacking Models based on Complementary Traffic Embeddings. In *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. 261–272. doi:10.1109/EuroSPW61312.2024.00035

[10] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*. PMLR, 1321–1330.

[11] Sayar Ul Hassan, Jameel Ahamed, and Khaleel Ahmad. 2022. Analytics of machine learning-based algorithms for text classification. *Sustainable Operations and Computers* 3 (2022), 238–248. doi:10.1016/j.susoc.2022.03.001

[12] Anh Ngo, Agri Candri, Teddy Ferdinan, Jan Kocon, and Wojciech Korczynski. 2022. StudEmo: A Non-aggregated Review Dataset for Personalized Emotion Recognition. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma (Eds.). European Language Resources Association, Marseille, France, 46–55. https://aclanthology.org/2022.nlperspectives-1.7/

[13] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* 45, 4 (2009), 427–437. doi:10.1016/j.ipm.2009.03.002