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

Samuel B. Jesus[1], **Guilherme Dal Bianco**[2],**Valerio Basile**[3], **Marcos André Gonçalves**[1]

[1]Department of Computer Science – Federal University of Minas Gerais

[2]Universidade Federal da Fronteira Sul - Campus Chapecó, Brazil

[3]Department of Computer Science – University of Turin, Turin, Italy

guilherme.dalbianco@uffs.edu.br, {samuelbrisio,mgoncalv}@dcc.ufmg.br, valerio.basile@unito.it

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.*

Resumo. *Em contextos subjetivos, como a detecção de discurso de ódio ou ironia, a classificação de textos é uma tarefa interpretativa que depende da bagagem cultural. Diferentemente de métodos tradicionais que agregam opiniões, por exemplo, por voto majoritário, o perspectivismo explora o conhecimento de grupos específicos de anotadores para construir modelos mais equitativos e representativos. No entanto, abordagens perspectivistas costumam demandar alto custo computacional, especialmente aquelas que recaem no ajuste fino de modelos de linguagem pré-treinados. Neste contexto, este trabalho tem dois objetivos principais: (i) investigar métodos tradicionais de aprendizado de máquina (como SVM, Random Forest e XGBoost) visando à redução de custos; e (ii) aplicar calibração para mitigar desequilíbrios na geração de inferências entre modelos perspectivistas. Os experimentos demonstraram ser possível reduzir o tempo de processamento em até 12 vezes , sem perda estatística na eficácia. Além disso, a calibração mostrou-se eficaz na redução de vieses de algumas perspectivas majoritárias, promovendo maior equidade entre os modelos.*

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 [Basile et al. 2021]. Uma característica inerente desse tipo de problema é o desacordo de rótulo (*label disagreement*) (ódio vs. não-ódio ou irônico vs.

não-irônico) [Aroyo and Welty 2015]. 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 [Fleisig et al. 2023]. A proposta do **perspectivismo** é preservar as múltiplas anotações para capturar diferentes visões, promovendo maior equidade [Frenda et al. 2024a]. 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 [Akhtar et al. 2021]. Particularmente, em [Casola et al. 2023], 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 [Casola et al. 2023] — doravante denominada método base — por meio da integração com modelos tradicionais de aprendizado de máquina, 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 [Frenda et al. 2024b]. Nessas situações, a divergência entre rótulos costuma ser tratada como ruído [Fleisig et al. 2023], adotando-se o voto da maioria e desconsiderando perspectivas de grupos minoritários potencialmente afetados [Akhtar et al. 2021]. 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 [Basile et al. 2021]. Tal abordagem tem ganhado destaque

[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.

diante da crescente demanda por modelos justos, inclusivos e sensíveis a vieses [Basile et al. 2021, Fleisig et al. 2023, Akhtar et al. 2021].

Em [Casola et al. 2023], 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. [Fleisig et al. 2023] 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 [Ngo et al. 2022] 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.

3. 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 [Casola et al. 2023], 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 [Guo et al. 2017]. 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

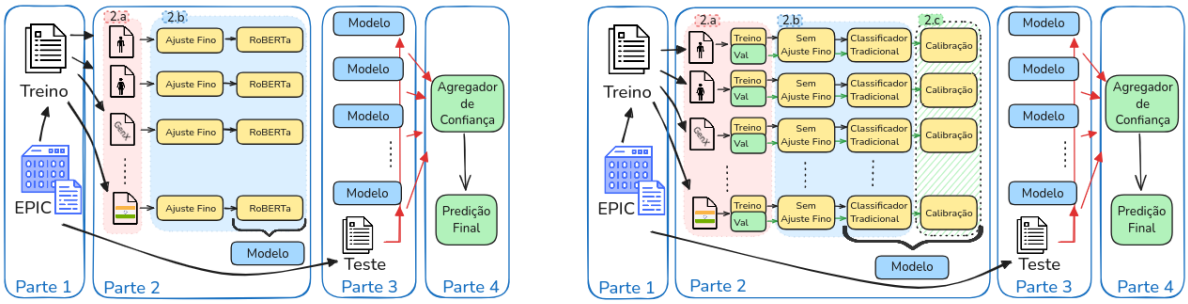

Figure 1. Método perspectivista original [Casola et al. 2023] ilustrado à esquerda e o método com as alterações propostas à direita.

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, 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.

4. 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.

4.1. Conjunto de dados

Para a avaliação experimental, foi utilizado o *English Perspectivist Irony Corpus* (EPIC) [Frenda et al. 2023]. 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.

4.2. Métricas de avaliação

Eficácia é mensurada pelo *macro F1-score* [Sokolova and Lapalme 2009], correspondente à média simples dos F1-scores por classe, atribuindo peso igual a todas. 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 cinco 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.

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

Table 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 *	$60.9 \pm 1.3 \downarrow$	63.1 ± 0.5 *
Somas das Confianças (SC)	67.2 ± 1.7	$\mathbf{65.2 \pm 1.0}$ *	$62.9 \pm 1.2 \downarrow$	64.3 ± 0.9 *
Voto Majoritário	67.0 ± 1.5	64.8 ± 1.5 *	$62.2 \pm 1.3 \downarrow$	64.4 ± 0.7 *
Sem-Perspectiva	65.1 ± 1.7	62.1 ± 1.6 *	$60.4 \pm 1.6 \downarrow$	$60.0 \pm 1.4 \downarrow$

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

4.3. 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 da de capturar relações lineares nos dados.

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 da 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 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

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

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. Sem calibração, as perspectivas "Ireland" e "GenX" concentram as maiores confianças, enquanto "GenY" e India" apresentam valores reduzidos. Com o método de agregação CM (que escolhe a predição de maior confiança), "Ireland" gera 48,9% das predições, e "India" não contribui,

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

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

evidenciando baixa equidade. 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 a objetivo original do perspectivismo.

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 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*.

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.

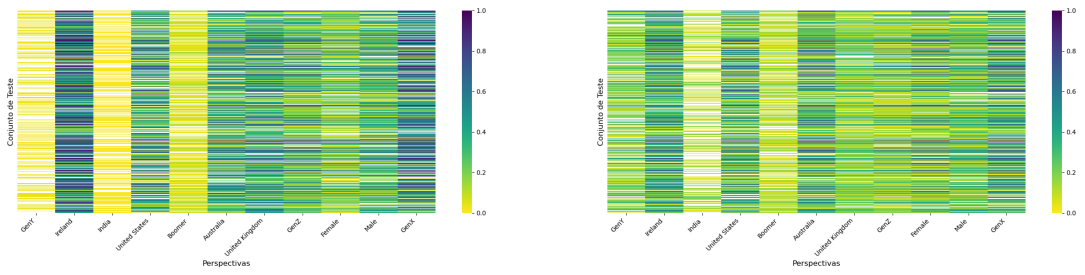

Figure 2. Mapa de calor com a confiança de cada perspectiva sem (esquerda) e com a calibração (direita). O tom azul representa alta confiança do modelo na perspectiva; amarelo, baixa confiança. Modelos com alta confiança (tom azul escuro) dominam a geração de inferências.

5. Contribuição

Nesta seção, são descritas as contribuições do aluno ao longo do desenvolvimento deste trabalho. As atividades foram organizadas em etapas, com o objetivo de apresentar de forma clara e estruturada o progresso alcançado e o papel desempenhado em cada fase do projeto.

A primeira etapa corresponde ao **Levantamento Bibliográfico**, no qual foi realizada uma pesquisa aprofundada sobre os seguintes temas: perspectivismo, desacordo de rótulos, calibração de modelos e modelos de linguagem de grande porte (LLMs). Dentre

esses, o *perspectivismo* foi o foco principal da investigação teórica, por ser o eixo central da proposta metodológica.

Na etapa seguinte, referente à **Implementação dos Modelos e Ajuste de Hiperparâmetros**, foi conduzido um estudo detalhado do código-fonte disponibilizado por [Casola et al. 2023], com o intuito de compreender sua estrutura e organização interna. A partir dessa análise, foram implementadas modificações nos trechos relacionados aos modelos de linguagem, de forma a permitir a aplicação do método de calibração conhecido como *Platt Calibration*. Além disso, no contexto das abordagens tradicionais utilizadas na proposta, foi realizada a implementação dos algoritmos correspondentes, bem como a integração com a base de código existente. Também foi conduzida uma busca sistemática por hiperparâmetros adequados, visando maximizar o desempenho dos modelos nessas configurações.

A etapa de **Coleta dos Resultados** consistiu na adaptação e modificação do código em pontos estratégicos, com o objetivo de garantir a correta extração das informações geradas durante os experimentos. Para isso, foi necessário planejar e organizar o armazenamento dos dados, a fim de facilitar a etapa posterior de análise.

Por fim, na etapa de **Análise dos Dados e Resultados**, foram desenvolvidos programas específicos para processar e resumir os dados coletados. No experimento relacionado à configuração *modelo* $\pm$ *calibração*, foram realizadas 11 execuções, variando-se a semente aleatória (*seed*) de 10 até 20, o que resultou em um volume expressivo de dados.

Com base nos resultados obtidos e em discussões realizadas com o orientador, foram identificadas limitações na abordagem proposta, bem como oportunidades de aprimoramento. Foram também elaborados gráficos e tabelas que possibilitaram a visualização e interpretação clara dos dados experimentais.

6. 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 [Casola et al. 2023] 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. Como trabalho futuro, planeja-se explorar outros conjuntos de dados perspectivistas e investigar técnicas de empilhamento [Gioacchini et al. 2024] visando maior efetividade com menor custo.

References

Akhtar, S., Basile, V., and Patti, V. (2021). Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *CoRR*, abs/2106.15896.

Aroyo, L. and Welty, C. (2015). Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *ACM Web Science 2013*.

Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M., and Uma, A. (2021). We need to consider disagreement in evaluation. In Church, K., Liberman, M., and Kordoni, V., editors, *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Casola, S., Lo, S. M., Basile, V., Frenda, S., Cignarella, A. T., Patti, V., and Bosco, C. (2023). Confidence-based ensembling of perspective-aware models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3496–3507, Singapore. Association for Computational Linguistics.

Fleisig, E., Abebe, R., and Klein, D. (2023). When the majority is wrong: Modeling annotator disagreement for subjective tasks. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Frenda, S., Abercrombie, G., Basile, V., Pedrani, A., Panizzon, R., Cignarella, A. T., Marco, C., and Bernardi, D. (2024a). Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, pages 1–28.

Frenda, S., Gavin Abercrombie, Basile, V., Pedrani, A., Panizzon, R., Cignarella, A. T., Marco, C., and Bernardi, D. (2024b). Perspectivist approaches to natural language processing: A survey. *Language Resources and Evaluation*.

Frenda, S., Pedrani, A., Basile, V., Lo, S. M., Cignarella, A. T., Panizzon, R., Marco, C., Scarlini, B., Patti, V., Bosco, C., and Bernardi, D. (2023). EPIC: Multi-perspective annotation of a corpus of irony. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.

Gioacchini, L., Santos, W., Lopes, B., Drago, I., Mellia, M., Almeida, J. M., and Gonçalves, M. A. (2024). Explainable stacking models based on complementary traffic embeddings. In *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 261–272.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Ngo, A., Candri, A., Ferdinan, T., Kocon, J., and Korczynski, W. (2022). StudEmo: A non-aggregated review dataset for personalized emotion recognition. In Abercrombie, G., Basile, V., Tonelli, S., Rieser, V., and Uma, A., editors, *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 46–55, Marseille, France. European Language Resources Association.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.