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

ANTHONY IRLAN MARQUES LUZ

APRIMORANDO A TAREFA DE DETECÇÃO DE IRONIA POR MEIO DE COMITÊS DE CLASSIFICADORES

PICOS - PI

2024

ANTHONY IRLAN MARQUES LUZ

APRIMORANDO A TAREFA DE DETECÇÃO DE IRONIA POR MEIO DE COMITÊS DE CLASSIFICADORES

Trabalho de Conclusão de Curso 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. Rafael Torres Anchieta.
Coorientador: Prof. Esp. Manoel Messias Pereira Medeiros

PICOS - PI

2024

APRIMORANDO A TAREFA DE DETECÇÃO DE IRONIA POR MEIO DE COMITÊS DE CLASSIFICADORES

Anthony Irlan Marques Luz
Rafael Torres Anchiêta
Manoel Messias Pereira Medeiros

RESUMO

Este artigo apresenta uma abordagem baseada em comitê de classificadores para identificar tweets irônicos escritos em português. A ironia é um fenômeno linguístico que pode ser percebido como um aspecto engraçado ou estranho de uma situação, utilizando palavras que dizem o oposto do que realmente significam, frequentemente como uma piada, e com uma entonação que demonstra isso. Quando se trata apenas de texto, detectar a ironia se torna bastante desafiador. Este trabalho avaliou diversas estratégias para detecção de tweets irônicos utilizando comitês de classificadores no corpus de tweets da tarefa compartilhada IDPT-21. O melhor resultado alcançado foi de 0,59 na acurácia balanceada utilizando a estratégia de *bagging*, superando o resultado estado da arte que era de 0,55, indicando que o comitê de classificadores auxiliou na melhora dos resultados.

Palavras-chave: Processamento de Língua Natural; ironia; técnicas de amostragem; Machine learning; Comitê de classificadores.

ABSTRACT

This paper presents an approach based on an ensemble to identify ironic tweets written in Portuguese. Irony is a linguistic phenomenon that can be perceived as a funny or peculiar aspect of a situation, using words that express the opposite of their literal meaning, often in a humorous context, and with intonation that reflects this. When dealing solely with text, detecting irony becomes quite challenging. This study evaluated several strategies for detecting ironic tweets using an ensemble in the tweet corpus of the shared task IDPT-21. The best result achieved 0.59 balanced accuracy using the bagging strategy, surpassing the state-of-the-art result of 0.55, indicating that the ensemble contributed to improving the results.

Keywords: Natural Language Processing; irony; sampling techniques; machine learning; ensemble.

Data de aprovação: XX/XX/XXXX (data de apresentação do TCC)

1 INTRODUÇÃO

Segundo o dicionário Oxford, a ironia pode ser definida como um aspecto engraçado ou estranho de uma situação muito diferente do que se espera, usando palavras que dizem o contrário do que realmente significam, muitas vezes como uma piada e com um tom de voz. Mas quando se trata apenas de texto, detectar a ironia torna-se uma tarefa ainda mais desafiadora, pois faltam outros elementos significativos para essa determinação. Por exemplo, na Figura 1 é apresentado um exemplo de um texto irônico.

Jogador do Inter apelidado de Valdívia
faz um golaço contra o Palmeiras

Figura 1. Exemplo de um texto irônico.

A ironia nessa frase é que um jogador apelidado de "Valdivia" (possivelmente em homenagem ao jogador chileno que jogou pelo Palmeiras) marcou um "brilhante gol" contra o time do Palmeiras, assim, um jogador com o mesmo apelido do ex-jogador do time adversário marcou um gol extraordinário contra a mesma equipe. Isso pode ser visto como uma reviravolta inesperada e irônica. Com esse texto, é possível perceber que identificar ironia sem contexto pode ser uma tarefa muito difícil, mesmo para humanos.

O interesse crescente no estudo da ironia levou a uma maior eficácia dos métodos de detecção de textos irônicos. Competições colaborativas entre pesquisadores desempenharam um papel crucial no avanço desses métodos em diversos idiomas. Por exemplo, o SemEval-2018(inglês), IroSvA-2019 (espanhol), WANLP-2021(árabe) e IDPT-21(português), esses foram eventos onde equipes de diferentes partes do mundo competiram para desenvolver os melhores modelos e algoritmos para a detecção de ironia em seus respectivos idiomas.

Especificamente, no caso do IDPT-21, que direcionou seus esforços para a identificação de textos irônicos em português, os resultados obtidos pelas equipes não alcançaram o patamar das outras tarefas compartilhadas. No IDPT-21 o desempenho mais destacado foi apenas de 0,52, enquanto que nas demais, as equipes participantes conseguiram atingir uma precisão superior. Posteriormente, esse resultado foi superado por Luz et. al. (2023), alcançando um valor de 0,55 (acc), adotando uma abordagem de balanceamento e seleção de *features*. Isso indica a existência de margem para aprimoramentos na área.

Neste artigo, dedicou-se atenção à otimização dos resultados obtidos anteriormente na detecção de ironia, explorando o potencial dos comitês de classificadores. Os Comitês têm se destacado como uma abordagem eficaz para melhorar o desempenho em tarefas complexas (Lima et al., 2019). Ao combinar as previsões de vários classificadores, é possível compensar as limitações individuais de cada modelo, proporcionando uma visão mais abrangente e robusta da tarefa em questão (Rahman et al., 2014).

Cinco conjuntos distintos foram treinados utilizando técnicas diversas, incluindo "*gradient boosting*", "*random forest*", "*bagging*", "*voting*" e "*stacking*". Esses *ensembles* foram aplicados à mesma tarefa e avaliados em três variações do mesmo conjunto de dados (*IDPT dataset*). Posteriormente, os resultados obtidos foram coletados e comparados com as performances anteriormente registradas por um único classificador considerado “fraco”.

Após as análises, constatou-se que o método de *Bagging* alcançou o melhor resultado, apresentando uma acurácia balanceada de 0,59 no *corpus* do IDPT-21. Este desempenho superou o resultado obtido no estudo anterior, que registrava uma taxa de 0,55. Esses resultados destacam o potencial dos comitês de classificadores em aprimorar as performances quando comparados a um único classificador isolado.

O restante do artigo está organizado da seguinte maneira. A seção 2 apresenta de maneira resumida os trabalhos relacionados, na seção 3, é apresentado o referencial teórico

para melhor compreensão do trabalho, na seção 4, a metodologia do trabalho é detalhada, a seção 5 faz uma análise dos resultados alcançados e por fim, a seção 6 conclui o trabalho e apresenta as considerações finais.

2 TRABALHOS RELACIONADOS

(Anchiêta et al. 2021) desenvolveram uma abordagem baseada em atributos superficiais, como a frequência do termo–inverso da frequência nos documentos (TF-IDF), e treinaram o classificador *Support Vector Machine* (SVM) para identificar textos irônicos. Além disso, os autores utilizaram a retro tradução como aumento de dados para equilibrar o corpus. O método alcançou uma precisão balanceada de 0,523.

(Oliveira et al. 2021) também adotaram o TF-IDF como um atributo e avaliaram três classificadores diferentes: Regressão Logística, *Naive Bayes* e *Random Forest*. Os autores atingiram uma precisão balanceada de 0,511 com o classificador *Naive Bayes*.

(Heinrich et al. 2021) exploraram diferentes estratégias de pré-processamento, extração de atributos e testaram dez algoritmos de aprendizado de máquina. O melhor resultado foi alcançado com o Perceptron Multicamadas, juntamente com uma etapa de pré-processamento nos dados, e também a medida TF-IDF como característica. Esta abordagem alcançou uma precisão balanceada de 0,502.

(Luz et al., 2023) conceberam uma estratégia fundamentada na seleção de padrões distintivos e descritores frequentemente associados a textos irônicos. Essa abordagem, aliada à utilização do algoritmo de árvore de decisão para a classificação, culminou em uma acurácia balanceada de 0,55.

3 REFERENCIAL TEÓRICO

3.1 Processamento de Língua Natural

O processamento de língua natural (PLN), também denominado Linguística Computacional ou, ainda, Processamento de Línguas Naturais, constitui um subcampo da ciência da computação dedicado à aplicação de técnicas computacionais para a aprendizagem, compreensão e produção de conteúdo em linguagem humana (Hirschberg e Manning, 2015).

PLN engloba a criação de modelos computacionais voltados para a execução de diversas tarefas que requerem a compreensão e manipulação de informações expressas em línguas naturais. Isso inclui atividades como tradução e interpretação de textos, busca por informações em documentos e interação homem-máquina (Russell; Norvig, 2016).

Para que estas atividades sejam executadas de maneira eficiente, tarefas de PLN requerem acesso a uma vasta base de conhecimento, originada de uma ampla coleção de textos especializados no domínio em questão (Cambria; White, 2014). Esse conjunto é geralmente chamado de corpus.

Dessa forma, aplicações específicas de PLN, demandam uma robusta base de dados textuais, contendo uma grande quantidade de textos de referência no idioma e domínio específicos da aplicação em questão.

3.2 Aprendizado de Máquina

A Aprendizagem de Máquina (AM) confere aos sistemas a capacidade de aprender e aprimorar-se automaticamente, fundamentando-se na experiência adquirida por meio de treinamentos com dados. Tanto no processo de treinamento quanto na fase de predição, os métodos de AM usualmente adotam uma das seguintes abordagens: supervisionada, não

supervisionada, por reforço ou semi-supervisionada. Independentemente da abordagem escolhida, o objetivo permanece em gerar previsões, classificações e/ou representações a partir de um conjunto de características de entrada (Jordan; Mitchell, 2015).

Na abordagem supervisionada (adotada neste trabalho), os dados de treinamento são representados como uma coleção de pares (x, y), onde o objetivo é gerar uma previsão *y* (por exemplo, “irônico ou não irônico”, “SPAM ou não SPAM”, etc.) em resposta a uma consulta *x* (como textos, documentos, etc.) (Jordan; Mitchell, 2015).

Tarefas de aprendizado supervisionado ainda podem ser divididas em mais dois grupos, classificação e regressão. Embora ambos objetivem, de maneira geral, mapear entradas para saídas, os problemas de classificação e regressão apresentam diferenças substanciais. Enquanto o primeiro gera um rótulo de classe discreta como valor de saída, como, por exemplo, "irônico" ou "não irônico", o segundo concentra-se em mapear entradas para valores contínuos reais de saída, como o preço de um produto (Hastie; Tibshirani; Friedman, 2009).

3.3 Comitês de classificadores

Como mencionado anteriormente, a tarefa de classificação envolve a distinção de um objeto específico entre duas ou mais classes, embora essa prática possa apresentar imprecisões. De acordo com Rahman et al. (2024), dependendo da distribuição dos padrões, é possível que nem todos os padrões sejam bem aprendidos por um classificador individual. Uma forma de resolver esse problema é usando conjuntos de classificadores para a mesma tarefa prática, que recebe o nome de *“ensemble classifier”* em português, comitês de classificadores (Polikar; Robi, 2006).

Um comitê de classificadores opera da seguinte maneira: diversos classificadores são treinados de forma individual, e seus resultados são armazenados. Em seguida, é aplicada uma técnica de combinação nesses resultados, determinando assim o produto final do comitê, conforme apresentado na Figura 2 (Rahman et al. 2024).

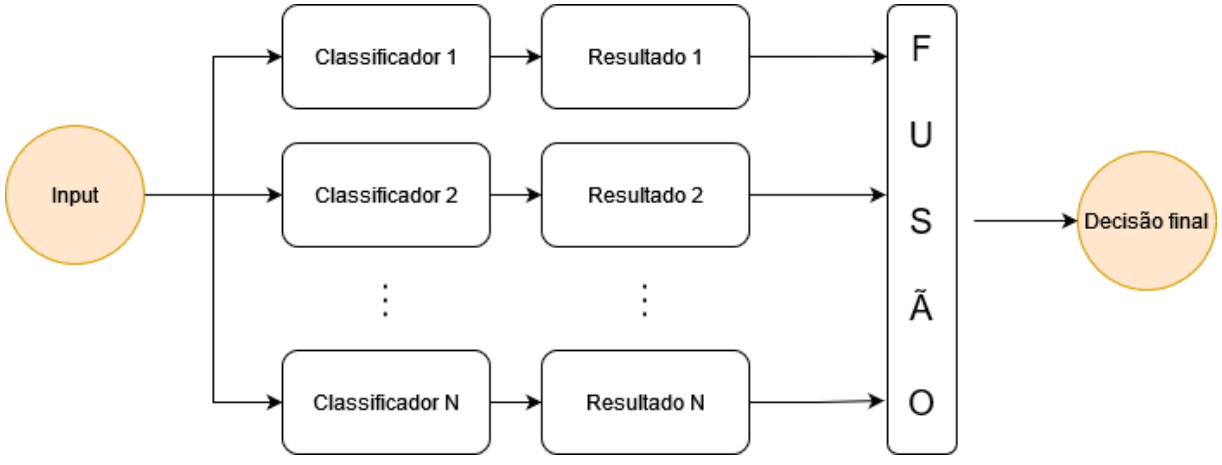

Figura 2. Exemplo do funcionamento de um comitê de classificadores.

Existem diversas estratégias para realizar a fusão desses resultados, tais como: *Gradient boosting*, *random forests*, *bagging*, *voting*, *stacking,* entre outras[1].

[1] https://scikit-learn.org/stable/modules/ensemble.html

4 METODOLOGIA

Visando otimizar os resultados previamente obtidos, foram conduzidos testes em três diferentes cenários nas bases de dados, a saber:

1. **Base de dados original.**

   A base de dados do IDPT-21 contém 12.736 *tweets* irônicos e 2.476 *tweets* não irônicos.

2. **Base de dados aumentada.**

   Nesta versão expandida, os 2.476 tweets não irônicos da base original foram duplicados de forma aleatória, mantendo, assim, a proporção de 12.736 tweets irônicos e não irônicos.

3. **Base de dados diminuída**

   Na versão reduzida, dos 12.736 tweets irônicos da base original, 10.260 foram selecionados e removidos de maneira aleatória, preservando a proporção de 2.476 tweets irônicos e não irônicos.

   Na Figura 3, é ilustrado as distribuições utilizadas neste trabalho.

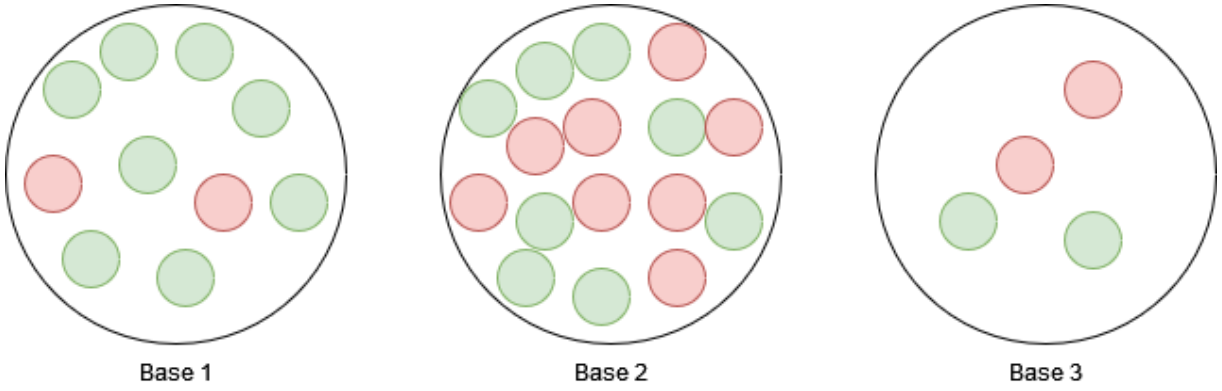

Figura 3. Exemplo visual da distribuição das bases de dados, as instâncias em verde representam *tweets* irônicos enquanto as em vermelho os *tweets* não irônicos.

Os cenários de bases de dados originais, aumentada e diminuída, proporcionaram uma visão abrangente sobre a capacidade dos modelos de lidar com diferentes quantidades de dados e distribuições. A base de dados original, composta por 12.736 tweets irônicos e 2.476 tweets não irônicos, representou o cenário inicial. A base aumentada, onde os tweets não irônicos foram duplicados mantendo a proporção original, buscou avaliar a influência da quantidade de dados não irônicos no desempenho dos modelos. Por fim, a base diminuída, na qual os tweets irônicos foram aleatoriamente removidos, investigou a resiliência dos modelos diante de uma redução de dados irônicos.

A partir desse ponto, foram utilizados os extratores de características textuais nesta base de dados. Esses extratores são compostos por padrões textuais identificados no corpo do tweet, tais como a presença de pontuação excessiva, o uso de expressões específicas, todos definidos no trabalho de Luz et. al. 2023.

Conforme ilustrado na Figura 4, a metodologia foi dividida em duas etapas. Na primeira etapa, cada comitê foi treinado em três iterações, correspondendo a cada uma das bases de dados. Em seguida, o classificador que obteve a maior acurácia balanceada foi aplicado a uma nova base de dados (também fornecida pelo IDPT). Isso permitiu a mensuração objetiva do seu desempenho, que foi posteriormente registrado para análises subsequentes.

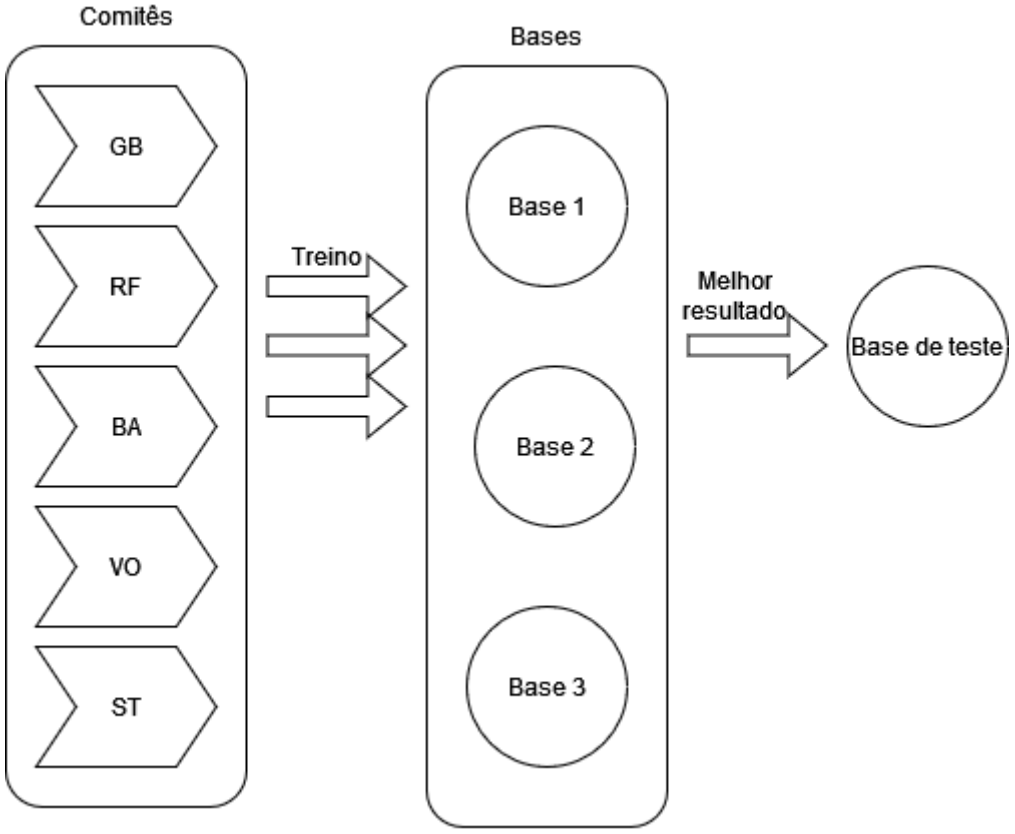

Figura 4. Metodologia adotada.

Para cada uma dessas configurações, foram empregados os respectivos comitês *Gradient boosting (GB)*, *random forests (RF)*, *bagging (BA)*, *voting (VO)*, *stacking (ST)*, todos utilizando os hyperparametros padrões de cada comitê, disponíveis na biblioteca "*scikit-learn*" e implementados na linguagem de programação *Python.*

5 ANÁLISE E RESULTADOS

Após avaliar as diversas configurações apresentadas na metodologia, o melhor resultado atingiu a pontuação de 0,59 de acurácia balanceada (medida utilizada como padrão entre os trabalhos do IDPT-21 para fins de comparação), usando o método de *Bagging,* na base de dados aumentada.

O *Bagging* é uma técnica de *ensemble* que busca produzir uma predição (seja para classificação ou regressão) por meio da agregação de resultados provenientes de diversos classificadores ou regressores “fracos”, utilizando o mecanismo de votação.

Iniciando o processo, o Bagging divide o conjunto de dados original em múltiplos subconjuntos e, subsequentemente, treina um único algoritmo em cada uma dessas subdivisões. Posteriormente, esses resultados são combinados, empregando votação no caso

de problemas de classificação e média para problemas de regressão, convergindo assim em uma única predição consolidada[2].

No caso do presente trabalho, o algoritmo utilizado no treinamento do comitê foi o KNN (do inglês, *k-Nearest Neighbors*), como o nome sugere, o KNN faz a classificação do dado com base na distância dos vizinhos de mesmo rótulo[3].

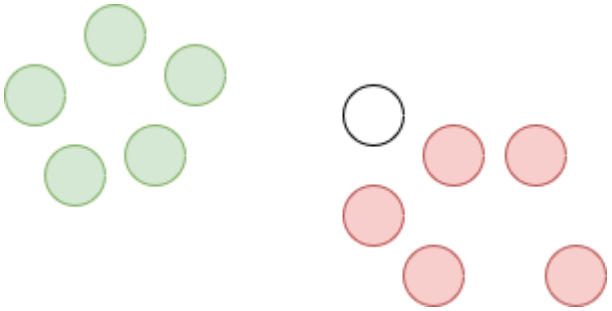

Figura 5. Exemplificação de como o KNN classifica novas instâncias.

Como mostra a Figura 5, as amostras em verde e vermelho representam os dados já presentes no conjunto de treinamento. Após a introdução de um novo dado, o modelo avalia a classe com base nas instâncias armazenadas mais próximas. No exemplo apresentado, o novo dado (representado pelo círculo branco) é classificado como pertencente à classe vermelha, determinado pela maioria das instâncias mais próximas, evidenciadas pelas amostras vermelhas na vizinhança do ponto.

Na Tabela 1, são apresentados os resultados de todos os comitês de classificadores.

Tabela 1. Resultados de todos os métodos.

Método	Base original	Base aumentada	Base diminuída
Gradient Boosting	0,50	0,55	0,54
Random Forest	0,50	0,55	0,55
Bagging	0,50	**0,59**	0,55
Voting	0,50	0,55	0,57
Stacking	0,50	0,54	0,56

Já na Tabela 2, são apresentados os resultados de outros grupos de pesquisa utilizando a mesma base de dados de tweets.

[2]https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html%23rb1846455d0e5-1&sa=D&source=docs&ust=1706014739123784&usg=AOvVaw1M0WyCi6GhgjM6uHQzDU9f

[3]https://scikit-learn.org/stable/modules/neighbors.html&sa=D&source=docs&ust=1706014739124454&usg=AOvVaw2Q84ebwunXyGJYa7xCE4w8

Tabela 2. Resultados de outros trabalhos.

Trabalhos	**ACC**
Método Proposto	0,59
(Luz et al. 2023)	0,55
(Anchiêta et al. 2021)	0,52
(Oliveira et al. 2021)	0,51

Embora o aumento observado não tenha atingido níveis expressivos, o emprego de comitês de classificadores revelou-se benéfico quando contrastado com o desempenho de um único classificador.

Para mais informações sobre a execução, consulte o código disponível em: *https://github.com/Anth0nYM/comites-classificadores*

6 CONCLUSÃO

Este artigo forneceu uma análise comparativa de diversos comitês de classificadores aplicados à detecção de ironia em tweets redigidos em língua portuguesa. Após as investigações, constatou-se que o método de *Bagging* destacou-se ao alcançar os resultados mais promissores, evidenciando melhorias em relação às avaliações anteriores. Além disso, observou-se que os desempenhos em conjuntos de dados balanceados, com uma distribuição uniforme de 50% de presença para cada classe, superaram aqueles em conjuntos desbalanceados.

Para futuras pesquisas, almeja-se investigar técnicas de geração de texto com o intuito de expandir ou modificar a base de dados, integrando diversas manifestações de expressão irônica, sarcástica ou ambígua.

Adicionalmente, planeja-se submeter os modelos a testes em situações do mundo real. Essa avaliação prática permitirá a análise de sua capacidade de enfrentar desafios inerentes a ambientes dinâmicos e complexos.

REFERÊNCIAS

CORRÊA, Ulisses B. et al. Overview of the idpt task on irony detection in portuguese at iberlef 2021. Procesamiento del Lenguaje Natural, v. 67, p. 269-276, 2021.

LUZ, Anthony IM et al. Improving Irony Detection by Balancing Methods and Feature Selection. In: Anais do XII Brazilian Workshop on Social Network Analysis and Mining. SBC, 2023. p. 216-221.

ANCHIÊTA, Rafael T. et al. PiLN IDPT 2021: Irony Detection in Portuguese Texts with Superficial Features and Embeddings. In: IberLEF@ SEPLN. 2021. p. 917-924

OLIVEIRA, Hugo Gonzalo; PEREIRA, José; CRUZ, Guilherme. CISUC at IDPT2021: Traditional and Deep Learning for Irony Detection in Portuguese. In: IberLEF@ SEPLN. 2021. p. 898-909.

HEINRICH, Tiago; CESCHIN, Fabrício; MARCHI, Felipe. TeamUFPR at IDPT 2021: Equalizing a Strategy Using Machine Learning for Two Types of Data in Detecting Irony. In: IberLEF@ SEPLN. 2021. p. 925-932

HIRSCHBERG, Julia; MANNING, Christopher D. Advances in natural language processing. Science, v. 349, n. 6245, p. 261-266, 2015.

RUSSELL, S.; NORVIG, P. Artificial Intelligence: A Modern Approach. Pearson, 2016. ISBN: 9781292153964. Disponível em: https://books.google.com.br/books?id=XS9CjwEACAAJ.

CAMBRIA, E; WHITE, B. "Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]," in IEEE Computational Intelligence Magazine, vol. 9, no. 2, pp. 48-57, May 2014, doi: 10.1109/MCI.2014.2307227.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. Science, v. 349, n. 6245, p. 255–260, 2015. Disponível em: https://www.science.org/doi/abs/10.1126/science.aaa8415.

BARATTO, Gabriel Junges. Comparação entre os modelos pré-treinados GPT-3 e BERT na estimativa de esforço de software por analogia a partir de requisitos textuais. 2022. Trabalho de Conclusão de Curso. Universidade Tecnológica Federal do Paraná

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer New York, 2009. (Springer Series in Statistics). ISBN 9780387848587. Disponível em: https://doi.org/10.1007/978-0-387-84858-7.

LI, Junyi et al. Textbox: A unified, modularized, and extensible framework for text generation. arXiv preprint arXiv:2101.02046, 2021

YANG, Zhen et al. CSP: code-switching pre-training for neural machine translation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020. p. 2624-2636.

GUAN, Wang; SMETANNIKOV, Ivan; TIANXING, Man. Survey on automatic text summarization and transformer models applicability. In: Proceedings of the 2020 1st International Conference on Control, Robotics and Intelligent System. 2020. p. 176-184.

DOS SANTOS LIMA, Adriana et al. Classificador Ensemble: Uma Abordagem Não Paramétrica Aplicado à Detecção De Diabetes. Revista do Seminário Internacional de Estatística com R, v. 4, n. 2, 2019

RAHMAN, Akhlaqur; TASNIM, Sumaira. Ensemble classifiers and their applications: a review. arXiv preprint arXiv:1404.4088, 2014

POLIKAR, Robi. Ensemble based systems in decision making. IEEE Circuits and systems magazine, v. 6, n. 3, p. 21-45, 2006.