Comparativo de Algoritmos de Aprendizado de Máquina para a Classificação de Notícias sobre a Politec em Mato Grosso

Thiago Ruiz Lobo[1], Claudia Aparecida Martins[1]

[1]Instituto de Computação – Universidade Federal de Mato Grosso (UFMT), Av. Fernando Correa da Costa 2367, Boa Esperança, 78.060-900 – Cuiabá – MT – Brasil

thiago.ruiz.trl@gmail.com, claudia@ic.ufmt.br

Abstract. *This work applied five machine learning algorithms to classify and evaluate news headlines about Politec de Mato Grosso. For each technique used, a comparison was made using two feature extraction tools (BoW and TF-IDF) and three class balancing methods (Random Oversampling, SMOTE and SMOTE + Tomek Links). The results obtained show the efficiency of the class balancing methods and among the five machine learning techniques, the Multinomial Naive Bayes technique stands out, which obtained the best headline accuracy performance in a set of news items to which the models had no prior access.*

Resumo. *Este trabalho teve como objetivo a aplicação de cinco algoritmos de aprendizado de máquina para classificar e avaliar manchetes de notícias sobre a Politec de Mato Grosso. Para cada técnica utilizada foi feito um comparativo usando duas ferramentas de extração de características (BoW e TF-IDF) e três métodos de balanceamento de classes (Random Oversampling, SMOTE e SMOTE + Tomek Links). Os resultados obtidos mostram a eficiência dos métodos de balanceamento de classes e dentre as cinco técnicas de aprendizado de máquina, destaca-se a técnica de Multinomial Naive Bayes que obteve os melhores desempenhos de acurácia de manchetes em um conjunto de notícias que os modelos não tiveram acesso prévio.*

1. Introdução

Atualmente, o crescimento da área de Processamento de Linguagem Natural (PLN), juntamente com Inteligência Artificial voltado ao processamento e interpretação de textos, possibilita tarefas relacionadas à extração de informações de textos dos mais diversos, a identificação de entidades, como locais ou organizações, e também a categorização de acordo com o conteúdo ou de acordo com a emoção que um determinado texto expressa, classificando-o em positivo, neutro ou negativo. Esta última tarefa, conhecida como Análise de Sentimentos, segundo Prasad *et al.* (2023) pode ser aplicada em diversos domínios, possibilitando que empresas ou instituições embasam decisões a partir da avaliação de sentimentos, ou opiniões, presentes em redes sociais ou sites de notícias.

A Perícia Oficial e Identificação Técnica (Politec), presente em 18 unidades espalhadas pelo estado, realiza incontáveis quantidades mensais de laudos periciais desempenhando um papel crucial na elucidação de crimes. Ademais, a Politec é responsável pelas emissões de carteiras de identidade nas mais de 140 unidades de

identificação espalhadas pelo estado. Como consequência da grande quantidade de laudos e solicitações de serviço, por vezes, a ânsia por respostas gera repercussões negativas em sites de notícia que carregam a imagem da instituição para a sociedade. Logo, entender como as manchetes de notícias impactam na imagem da Politec perante a sociedade é um desafio valoroso que pode ser analisado e utilizado para melhor divulgar o trabalho realizado. Portanto, neste trabalho foram realizados vários experimentos com o objetivo de comparar o desempenho das técnicas de aprendizado de máquina na análise de sentimentos em manchetes de notícias sobre a Politec divulgadas nas principais mídias de Mato Grosso. Dessa forma, essa análise permitirá que a instituição possa ter a percepção das principais notícias veiculadas sobre si no estado pelos veículos locais.

O artigo foi dividido em cinco seções, sendo a Seção 2 a revisão de literatura na qual são apresentados os trabalhos relacionados que aplicaram técnicas de aprendizado de máquina, além dos principais conceitos das técnicas utilizadas no desenvolvimento deste trabalho. Na Seção 3 é descrita a metodologia utilizada na classificação de manchetes de notícias sobre a Politec. Na Seção 4 são apresentados e analisados o desempenho dos resultados obtidos e, por fim, na Seção 5 a conclusão e trabalhos futuros dos principais aspectos identificados no desenvolvido do trabalho.

2. Revisão de Literatura

2.1. Trabalhos relacionados

Diversas pesquisas dedicam-se ao estudo e aplicação de técnicas de aprendizado de máquina e análise de sentimentos de sites de notícias ou redes sociais, buscando identificar a polaridade do conteúdo ou, até mesmo, a veracidade do texto.

Nesse contexto, Maada *et al.* (2022) aplica uma abordagem de aprendizado de máquina no campo de análise de sentimentos, por meio do algoritmo de *Support Vector Machine* (SVM), combinado as técnicas de extração de características TF-IDF e N-gram em duas bases de dados, sentiment140 e Amazon reviews.

Em Jariwala *et al.* (2020) são comparados os resultados do uso das técnicas de SVM, *K-means* Clusterizado e *Naive Bayes* na análise de sentimentos de manchetes de notícias sobre o mercado de ações de duas empresas (A e B) com o intuito de predizer os movimentos do dia subsequente, a partir da polaridade das manchetes.

Já em Anitha e Gnanasekaran (2023) são comparadas as técnicas de SVM, *Multinomial Naive Bayes*, Regressão Logística e Árvore de Decisão aplicadas na área de análise de sentimentos para a classificação de tweets, mostrando que, para o domínio utilizado, o algoritmo de Regressão Logística teve o melhor desempenho comparado às outras três abordagens devido à sua força computacional e simplicidade.

2.2. Materiais e métodos

2.2.1 *Web-scraping* e pré-processamento

Web scraping, ou raspagem de dados, é um processo para coletar dados de sites de forma automatizada usando scripts. Pode ser aplicado em diferentes contextos, desde a coleta de preços de sites de lojas de varejo, como também na coleta de informações ilícitas presentes por vezes na darknet que podem ser fornecidas a autoridades legais.

Após a coleta dos dados, é necessário realizar o pré-processamento para o tratamento das informações. Nesta etapa, geralmente, é realizada a tokenização, no qual, um texto é dividido em unidades menores, normalmente, em palavras (*tokens*). Após, é realizada a eliminação de caracteres especiais e pontuações, a normalização de palavras maiúsculas em minúsculas e a eliminação de palavras comuns, presentes em uma lista denominada *Stopwords*, que são pouco relevantes na discriminação de um domínio. Também, pode-se aplicar outras duas técnicas: a stemização e a lematização. A técnica de stemização consiste na remoção de sufixos e/ou prefixos da palavra, o que poderá fazer com que palavras perdem seu significado. Já a técnica de lematização tem por objetivo encontrar o *lemma* em comum com outras palavras, sendo tal processo mais custoso computacionalmente, porém preserva o sentido original das palavras.

2.2.2 Extração de características

Segundo Maada *et al.* (2022), a extração de características é uma etapa vital na análise de sentimento, no qual aspectos relevantes dos dados são identificados para serem processados com técnicas de aprendizado de máquina. Técnicas como *Bag of Words* (BoW) e *Term Frequency - Inverse Distribution Frequency* (TF-IDF) são comumente utilizadas para extração de características, transformando palavras presentes em textos em vetores de números. O BoW representa a frequência das palavras presentes no texto, sem discriminar a relevância da palavra com todo o conjunto. Já a técnica TF-IDF utiliza duas medidas combinadas: a frequência de uma palavra (TF) no texto e, inversamente, a frequência da palavra em todo o conjunto de dados (IDF). Esta representação considera a importância de cada palavra e sua relevância no conjunto.

2.2.3 Balanceamento de classes

Para além da extração de características, na tarefa de classificação, muitos conjuntos de dados exigem o balanceamento de suas classes. De acordo com Wongvorachan *et al.* (2023), os métodos de reamostragem funcionam de diferentes formas, dividindo-os em três categorias: *Oversampling*, *Undersampling* e *Hybrid Sampling*. O *Oversampling* funciona replicando amostras da classe minoritária por meio de diversas abordagens, como a *Random Oversampling* (ROS), a qual replica a base minoritária de forma aleatória e a *Synthetic Minority Over-sampling Technique* (SMOTE), que gera novos exemplos da classe minoritária, identificando os exemplos mais próximos da classe minoritária no espaço de características. O *Under-sampling* funciona removendo amostras da classe majoritária e, assim como o *Oversampling*, possui as suas variações como o *Random Undersampling* (RUS) e o *Tomek's link Undersampling*. Por fim, existem técnicas que combinam os dois universos, chamadas de hybrid sampling. Algumas variações populares de reamostragem híbrida são a combinação de SMOTE com a técnica de *Tomek's Link*.

2.2.4 Técnicas de aprendizado de máquina

Vários algoritmos de aprendizado de máquina são utilizados no processamento de textos, utilizando estratégias probabilísticas para compreensões textuais. Awwalu *et al.* (2020) descreve o algoritmo *Multinomial Naive Bayes* (MNB) como uma das variações do Naive Bayes (NB), usualmente, utilizado para resolver problemas de classificação relacionados a textos, por meio do cálculo da probabilidade de um documento pertencer a uma determinada classe, a partir do conhecimento a priori, da presença de palavras

naquela classe. Silveira *et al.* (2021) cita que, Regressão Logística (RL) se diferencia da Regressão Linear porque a variável dependente é qualitativa e binária. Na RL a variável resposta assume apenas valores 0 ou 1, sendo geralmente "1" a ocorrência do evento de interesse e "0" a sua ausência. Já Yang (2022) apresenta que o algoritmo *Support Vector Machine* (SVM) inicialmente representa os pontos de dados como um grupo de vetores, determinando um limite desses vetores através de um hiperplano, separando assim em classes diferentes. Isso garante que vetores com características semelhantes àquelas próximas ao limite não são classificados na classe errada. Amer e Siddiqui (2020) detalham que o classificador de Árvore de Decisão (AD) refere-se a um simples modelo que representa decisões e seus possíveis resultados. Por fim, Jalal *et al.* (2022) mostra que o algoritmo de *Random Forest* (RF) utiliza um grande número de AD para a tomada de decisões, tomando a sua decisão final a partir da média dos resultados das árvores de decisão.

3. Metodologia

Nesta seção são apresentadas as etapas da metodologia utilizada, mostrada na Figura 1.

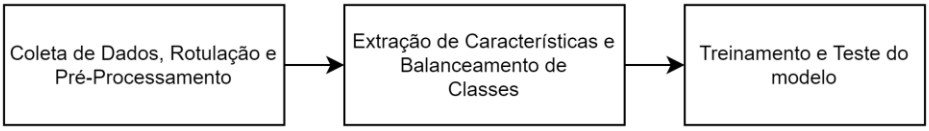

Figura 1. Etapas da metodologia

Para a coleta de dados textuais foi escolhido um conjunto de seis veículos de comunicação de notícias estaduais, evitando assim notícias sobre perícias de outros estados. A coleta de manchetes de notícias foi realizada usando algoritmos de web-scraping na linguagem Python e no ambiente virtual Google Colab. Para cada notícia, foi extraída a sua data de publicação, manchete e artigo da notícia. Ao final, um conjunto de 28572 notícias distribuídas nos anos de 2005 a 2024 foram armazenadas em um arquivo .csv.

Após, foram selecionadas notícias referentes ao ano de 2023 com 4051 amostras que foram rotuladas em notícias positivas, negativas ou neutras às ações da Politec. Todavia, devido ao alto grau de desbalanceamento entre as classes, foi necessário reduzir os dados a 595 amostras e uma melhor distribuição entre as classes.

Durante o pré-processamento, também foi realizado, para cada manchete de notícia, o processo de tokenização e a eliminação de sinais de pontuação e caracteres especiais. Em seguida, foi feita a normalização das letras maiusculas em minúsculas e a eliminação das stopwords. Por fim, foi aplicada a técnica de lematização.

Após o pré-processamento, foi realizada a etapa de extração das características, usando as técnicas BoW e TF-IDF criando, assim, dois conjuntos de dados distintos. Os conjuntos de dados foram divididos com uma proporção de 80% das manchetes para treino e 20% para teste. Para alcançar um melhor quantidade de amostras por classe, foram aplicadas técnicas de *Random Oversampling*, SMOTE e SMOTE + *Tomek Links*.

Por fim, na etapa de treinamento do modelo, foram aplicadas cinco técnicas de aprendizado de máquina, em conjunto às técnicas de extração de características, de balanceamento de classes e as técnicas de processamento RF, MNB, RL, SVM e AD.

Assim, para cada técnica de treinamento escolhida, foi feita uma combinação de ferramentas de extração de características e de balanceamento de classes, buscando posteriormente comparar os resultados e, assim, definir as melhores técnicas para classificação de manchetes de notícias.

4. Resultados

Na Tabela 1 são apresentados os resultados obtidos para o Conjunto Total de Notícias (CTN), com 4051 manchetes, e para o Conjunto Reduzido de Notícias (CRN), com 595 manchetes. Os valores apresentam as taxas de acertos aplicados em um novo conjunto de notícias.

Tabela 1. Resultados - CTN e CRN de 2023.

		BoW					**TF-IDF**				
		RF	**MNB**	**RL**	**SVM**	**AD**	**RF**	**MNB**	**RL**	**SVM**	**AD**
CTN	Original	0.96	0.92	0.96	0.95	0.95	0.96	0.94	0.95	0.94	0.95
	Random Oversampling	0.96	0.93	0.95	0.95	0.95	0.96	0.92	0.94	0.95	0.94
	SMOTE	0.90	0.92	0.93	0.94	0.84	0.95	0.92	0.94	0.95	0.94
	SMOTE + *Tomek Links*	0.64	0.85	0.62	0.83	0.60	0.95	0.90	0.93	0.95	0.90
CRN	Original	0.84	0.82	0.82	0.83	0.76	0.84	0.82	0.82	0.83	0.76
	Random Oversampling	0.83	0.85	0.85	0.84	0.82	0.83	0.85	0.85	0.84	0.82
	SMOTE	0.85	0.86	0.85	0.84	0.79	0.85	0.86	0.85	0.84	0.79
	SMOTE + *Tomek Links*	0.67	0.51	0.61	0.73	0.62	0.67	0.51	0.61	0.73	0.62

Apesar da melhor acurácia obtida nos modelos quando aplicados ao CTN, tais modelos não conseguiram identificar a maioria das notícias positivas e negativas, generalizando-as em notícias neutras, devido ao alto desbalanceamento das classes. Com a aplicação das técnicas de balanceamento de classes, houve uma melhora no acerto de notícias negativas e positivas, sendo que o modelo que mesclava o uso de MNB em conjunto ao TF-IDF e o *Random Oversampling* apresentou a melhor taxa de acerto quando aplicado em um conjunto de novas notícias (0.92). Os modelos aplicados ao CRN apresentaram uma acurácia menor comparado ao CTN. Todavia, vale ressaltar que os modelos identificaram melhor parte das manchetes de notícias positivas e negativas, uma vez que a quantidade de amostras de cada classe estava mais equilibrada. Com a aplicação das técnicas de balanceamento de classes, houve uma melhora no acerto de manchetes de notícias negativas e positivas. Dessa forma, a melhor taxa de acerto foi 0.85 e 0.86 para vários testes realizados, tanto usando BoW quanto TF-IDF. É possível verificar que a técnica de SMOTE + *Tomek Links* foi a que apresentou o menor desempenho nos modelos.

5. Conclusão e Trabalhos Futuros

Este trabalho teve o objetivo de processar, analisar e comparar os resultados de técnicas de aprendizado de máquina aplicadas na análise de sentimentos de manchetes de notícias referentes a Politec. Os resultados obtidos mostraram que mesmo em conjunto de dados inicialmente desbalanceados, o uso de técnicas como o *Random Oversampling* e SMOTE para o balanceamento possibilitou que os modelos desenvolvidos obtivessem uma melhor taxa de acertos. O uso de técnicas de aprendizado de máquina, em especial a *Multinomial Naive Bayes*, se mostrou adequado para o desafio proposto neste trabalho, uma vez que a complexidade de informações não era muito alta, consideradas as manchetes de notícias em diversos meios de comunicação. Por fim, como trabalho

futuro, destaca-se a necessidade de validar tais modelos nos artigos das manchetes de notícias e a construção de um processo de *web scraping* otimizado para extrair apenas as novas manchetes de notícias mês a mês.

6. Referências

Amer, A. and Siddiqui, T. (2020) "Detection of Covid-19 Fake News text data using Random Forest and Decision tree Classifiers", In: International Journal of Computer Science and Information Security.

Anitha, S. and Gnanasekaran, P. (2023) "Juncture of Text Preprocessing Techniques & Extracting Sentiment Analyzing of Micro-Blog Based on Machine LearningAlgorithms" In: International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)

Awwalu, J., Umar, N., Ibrahim, M. and Nonyelum, O. (2020) "A Multinomial Naive Bayes Decision Support System For Covid-19 Detection", In: FUDMA Journal of Sciences, p. 704-711.

Jalal, N., Mehmood, A., Choi, G. and Ashraf, I. (2022) "A novel improved random forest for text classification using feature ranking and optimal number of trees", In: Journal of King Saud University - Computer and Information Sciences, p. 2733-42.

Jariwala, G. Agarwal, H. and Jadhav, V. (2020). "Sentimental Analysis of News Headlines for Stock Market", In: IEEE International Conference for Innovation inTechnology.

Maada, L., Fararni, K., Aghoutane, B., Fattah, M. and Farhaqui, Y. (2022) "A comparative study of Sentiment Analysis Machine Learning Approaches", In: International Conference on Innovative Research in Applied Science.

Prasad, O., Nandi, S., Dogra, V. and Diwakar, D. (2023) "A systematic review of NLP methods for Sentiment classification of Online News Articles", In: International Conference on Computing Communication and Networking Technology.

Silveira, M., Barbosa, N., Peixoto, A., Xavier E. and Júnior, S. (2021) "Application of logistic regression in the analysis of risk factor associated with arterial hypertension", In: Research, Society and Development.

Yang, L. (2022) "A Brief Introduction of the Text Classification Methods", In: IEEE International Conference on Electrical Engineering, Big Data an Algorithms.

Wongvorachan, T., He, S. and Bulut, O. (2023) "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining", In: Multidisciplinary Digital Publishing Institute.