Análise do Classificador Ingênuo de Bayes para Detecção de Fake News

Eliezer Martins de Oliveira
Centro de Informática, Universidade Federal de Pernambuco
Recife, Brasil
emo3@cin.ufpe.br

Erick Vinicius Rebouças Cruz
Centro de Informática, Universidade Federal de Pernambuco
Recife, Brasil
evrc@cin.ufpe.br

Joana D'arc Oliveira do Nascimento
Centro de Informática, Universidade Federal de Pernambuco
Recife, Brasil
jdon@cin.ufpe.br

Paulo Vitor Barbosa Santana
Centro de Informática, Universidade Federal de Pernambuco
Recife, Brasil
pvbs@cin.ufpe.br

***Resumo*—This study investigates the effectiveness of the Naive Bayes classifier in detecting fake news. Using the Fake and Real News dataset, which consists of both legitimate and fraudulent articles, we assess the method's performance through textual classification metrics. The implementation leverages Python with Scikit-learn for modeling, NLTK for natural language processing, and visualization libraries for exploratory analysis.**

I. Introdução

A proliferação de fake news nas plataformas digitais representa um desafio global para a integridade da informação. Este trabalho avalia o classificador Naive Bayes, método probabilístico eficiente em problemas de alta dimensionalidade típicos de processamento de linguagem natural.

A estrutura do documento organiza-se como: Seção II apresenta os objetivos, Seção III discute a relevância do tema, Seção IV detalha a metodologia, Seção **??** o cronograma, e Seção VIII as referências.

II. Objetivos

- Avaliar a performance do Naive Bayes na classificação de textos como fake news
- Comparar variantes do algoritmo (MultinomialNB, BernoulliNB) em diferentes representações textuais
- Analisar métricas de desempenho como precisão, recall e F1-score

III. Justificativa

A detecção automatizada de fake news é crucial para combater a desinformação em escala. O Naive Bayes destaca-se por sua simplicidade e eficiência computacional, sendo particularmente adequado para análise de textos onde a independência entre palavras pode ser assumida como aproximação válida.

IV. Metodologia

Neste projeto, será abordada a detecção de fake news utilizando técnicas de aprendizado de máquina e análise exploratória de dados textuais. Inicialmente, será realizada uma análise para compreender a distribuição de termos linguísticos, frequência de palavras-chave e padrões lexicais característicos em notícias falsas. Posteriormente, será implementado o classificador Ingênuo de Bayes, técnica probabilística especialmente eficaz para processamento de linguagem natural devido à sua capacidade de lidar com alta dimensionalidade de dados textuais. Todo o código será desenvolvido em Python, empregando bibliotecas como Scikit-learn para modelagem e vetorização textual, Pandas para manipulação de dados, NLTK para pré-processamento linguístico, além de Matplotlib e Seaborn para visualização dos padrões textuais e resultados de classificação.

A. O Dataset

O dataset *Fake and Real News* contém:

- 44,898 artigos (23,481 fake e 21,417 reais)
- Estrutura típica: Título, conteúdo textual, assunto e rótulo
- Desafios: Variabilidade linguística e desbalanceamento de classes

B. Pré-processamento

Etapas críticas para análise textual:

1) Normalização (lowercase, remoção de stopwords)
2) Tokenização e lematização

C. Teorema de Bayes

O Teorema de Bayes é uma das ferramentas da probabilidade, ele oferece um caminho preciso para estimar a chance de ocorrência de fenômenos a partir de conhecimento pré-existente. Desenvolvido pelo matemático britânico Thomas Bayes (1702–1761), esse princípio vai além de fórmulas matemáticas; representa um sistema dinâmico para revisar premissas iniciais diante de descobertas recentes. o teorema é amplamente reconhecido devido à simplicidade conceitual, aliada a um potencial de uso que atravessa fronteiras disciplinares.

No aprendizado de máquina e na inteligência artificial, o Teorema de Bayes é frequentemente utilizado para desenvolver classificadores probabilísticos, como o Classificador Naive

Bayes, que calcula a probabilidade de uma instância pertencer a uma classe específica do conjunto de dados com base nos atributos apresentados. Apresenta aplicações em diversas outras áreas, como a medicina, onde auxilia no diagnóstico de doenças com base em sintomas observados, e engenharia, onde é utilizado para prever falhas em sistemas complexos, sejam estes mecânicos, elétricos, computacionais e dentre outros. Sua relevância se estende a áreas como economia, biologia e segurança da informação, tornando-se uma importante ferramenta para análise e tomada de decisões.

A essência desse método probabilístico é a interação entre o conhecido e o desconhecido. Por exemplo, ao avaliar a autenticidade de uma transação bancária: integrando os dados como localização geográfica, horários e valores discrepantes, o teorema permite quantificar o risco de fraude em tempo real. Outro exemplo prático surge na verificação de notícias, onde se mede a confiabilidade interpretando o histórico da fonte, consistência factual e análise de viés linguístico. Essa adaptabilidade explica sua adoção massiva em sistemas que exigem atualizações instantâneas, desde plataformas de comércio eletrônico até monitoramento de redes sociais.

Matematicamente, o Teorema de Bayes é expresso pela seguinte fórmula:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{1}$$

Onde:

- $P(A|B)$ é a probabilidade do evento A ocorrer, dado o evento B ocorreu.
- $P(B|A)$ é a probabilidade de evento B ocorrer, dado o evento A ocorreu.
- $P(A)$ e $P(B)$ são as probabilidades a priori de A e B, respectivamente.

Esta fórmula permite calcular probabilidades condicionais, sendo aplicada no Classificador Naive Bayes para estimar a probabilidade de uma instância pertencer a uma classe específica com base em suas características.

D. Classificador de Bayes

O classificador Naive Bayes é um modelo amplamente utilizado em aprendizado de máquina que se baseia no Teorema de Bayes para realizar classificações probabilísticas. O termo "Naive" (ingênuo, em inglês) refere-se à suposição simplificadora de que as variáveis preditoras no conjunto de dados são independentes entre si, ou seja, os atributos não possuem correlação. Apesar dessa suposição ser frequentemente irrealista, ela torna o modelo extremamente eficiente e prático para diversas aplicações.

Uma das principais razões para a popularidade do Naive Bayes é sua simplicidade e rapidez, que permitem uma implementação fácil e resultados precisos em cenários variados. Comparado a outros classificadores, seu desempenho se destaca em situações onde os dados são escassos ou altamente dimensionais, o que o torna particularmente adequado para tarefas envolvendo atributos discretos.

O Naive Bayes encontra aplicações em diversas áreas. Na análise de crédito, por exemplo, ele auxilia na avaliação de risco de clientes, ajudando instituições financeiras a tomar decisões informadas. Na análise de texto, o modelo é eficaz para identificar palavras-chave com base na frequência de termos, o que é especialmente útil na classificação de e-mails como spam ou não spam. Além disso, sua utilidade se estende ao campo médico, onde é empregado em diagnósticos, fornecendo suporte na identificação de condições de saúde com base em dados clínicos. Em engenharia, o modelo é frequentemente aplicado para detectar falhas em sistemas mecânicos, destacando-se por sua confiabilidade.

Neste projeto, o classificador Naive Bayes será implementado utilizando a biblioteca Scikit-learn, reconhecida por suas ferramentas de aprendizado de máquina. O modelo será empregado para analisar textos e contextos relacionados a notícias, classificando as instâncias de fake news com base em atributos relevantes, como a fonte e características textuais. Essa abordagem garantirá uma análise eficiente e um desempenho consistente, proporcionando insights valiosos para a detecção de notícias falsas e ajudando a combater a disseminação de desinformação.

E. Classificador Naive Bayes Multinomial

O classificador Naive Bayes Multinomial é um modelo probabilístico amplamente utilizado para a classificação de dados discretos, especialmente em tarefas de processamento de linguagem natural (PLN) e mineração de texto. Ele se baseia no Teorema de Bayes, que permite calcular a probabilidade de uma classe dado um conjunto de características, assumindo a independência condicional entre essas características.

Matematicamente, a probabilidade de uma classe dado um vetor de características é calculada como:

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)}. \tag{2}$$

No contexto do modelo multinomial, representa a frequência de ocorrência da característica (como uma palavra em um documento). A suposição de independência condicional implica que a probabilidade conjunta das características pode ser expressa como o produto das probabilidades individuais:

$$P(x|C_k) = \prod_{i=1}^{n} P(x_i|C_k). \tag{3}$$

A probabilidade condicional de uma característica dada a classe pode ser estimada por:

$$P(x_i|C_k) = \frac{count(x_i, C_k) + \alpha}{\sum_j count(x_j, C_k) + \alpha N}, \tag{4}$$

onde $count(x_i, C_k)$ representa a frequência da característica x_i na classe C_k, $\sum_j count(x_j, C_k)$ é a soma das contagens de todas as características que apareceram na classe C_k, N é o número total de características no vocabulário e α é o parâmetro de suavização de Laplace (tipicamente) para evitar probabilidades zero.

Na fase de inferência, um novo exemplo é classificado calculando a pontuação para cada classe e escolhendo aquela com maior valor:

$$\hat{C} = \arg\max_{C_k} \left(\log P(C_k) + \sum_{i=1}^{n} \log P(x_i|C_k) \right). \quad (5)$$

O Naive Bayes Multinomial é amplamente empregado em tarefas como classificação de documentos, análise de sentimentos, filtragem de spam e detecção de fake news, apresentando bons resultados, especialmente quando os dados possuem uma distribuição multinomial e as características representam contagens discretas.

F. Classificador Complement Naive Bayes

O classificador Complement Naive Bayes (CNB) é uma variante do Naive Bayes projetada para lidar com conjuntos de dados desbalanceados, onde certas classes possuem significativamente menos exemplos que outras. Essa abordagem é particularmente útil em tarefas de classificação textual, como detecção de spam ou categorização de tópicos com distribuição assimétrica. O CNB mantém a suposição de independência condicional entre características, mas calcula as probabilidades de forma complementar, mitigando o viés em direção às classes majoritárias.

Matematicamente, a probabilidade de uma classe C_k dado um vetor de características x segue o Teorema de Bayes:

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)}. \quad (6)$$

Contudo, ao invés de estimar $P(x|Ck)$ diretamente, o CNB calcula a probabilidade das características no complemento de Ck, ou seja, em todas as classes exceto Ck. A probabilidade complementar é dada por:

$$P(x|\tilde{C_k}) = \prod_{i=1}^{n} P(x_i|\tilde{C_k}), \quad (7)$$

onde $\overline{C_k}$ representa a união de todas as classes diferentes de C_k.

A estimativa da probabilidade condicional complementar para uma característica xi é suavizada com Laplace e definida como:

$$P(x_i|\tilde{C_k}) = \frac{\sum_{j \neq k} count(x_i, C_j) + \alpha}{\sum_{j \neq k} \sum_m count(x_m, C_j) + \alpha N}, \quad (8)$$

onde:

- $\sum_{j \neq k} count(x_i, C_j)$ é a frequência total de x_i em classes diferentes de C_k,
- $\sum_{j \neq k} \sum_m count(x_m, C_j)$ é a soma de todas as contagens de características nas classes complementares,
- N é o número total de características únicas no vocabulário,
- α é o parâmetro de suavização de Laplace.

Na fase de classificação, um novo exemplo é atribuído à classe que minimiza a evidência das características nas classes complementares:

$$\hat{C} = \arg\min_{C_k} \left(\log P(C_k) - \sum_{i=1}^{n} \log P(x_i|\tilde{C_k}) \right). \quad (9)$$

O Complement Naive Bayes é especialmente eficaz em cenários com desbalanceamento acentuado, como identificação de documentos raros, diagnóstico de anomalias ou categorização hierárquica de textos, onde a abordagem complementar reduz a influência de classes dominantes e melhora a sensibilidade a padrões minoritários.

G. Avaliação

Protocolo experimental:

- Divisão 80-20 para treino-teste
  - 80% dos dados são utilizados para treino, enquanto 20% são utilizados para teste.
- Validação cruzada estratificada (k=5)
  - Técnica usada para avaliar modelos de aprendizado de máquina, garantindo que cada divisão dos dados preserve a mesma proporção das classes da variável-alvo.
- Matriz de confusão
  - A matriz de confusão é uma tabela usada para avaliar o desempenho de um modelo de classificação, especialmente em problemas de classificação binária. Ela compara os valores previstos pelo modelo com os valores reais do conjunto de dados de teste. A matriz é estruturada da seguinte forma:

	Previsto Positivo	Previsto Negativo
Real Positivo	VP	FN
Real Negativo	FP	VN

A partir dela, calculamos diversas métricas importantes:

∗ Acurácia: Proporção de previsões corretas.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \quad (10)$$

∗ Precisão: Proporção de positivos previstos corretamente.

$$Precisão = \frac{VP}{VP + FP} \quad (11)$$

∗ Recall (sensibilidade): Habilidade do modelo encontrar os positivos reais.

$$Recall = \frac{VP}{VP + FN} \quad (12)$$

∗ F1-Score: Média Harmônica entre precisão e recall.

$$F1Score = 2 \cdot \frac{Precisão \cdot Recall}{Precisão + Recall} \quad (13)$$

- AUC-ROC
  - A curva ROC é um gráfico que mostra a relação entre duas métricas importantes para um classificador:
    * Recall (o mesmo da matriz de confusão)
    * Taxa de Falsos Positivos: Mede a proporção de exemplos negativos que foram classificados incorretamente como positivos.

$$Taxa = \frac{FP}{FP + VN} \tag{14}$$

  - A AUC-ROC é um número entre 0 e 1 que representa a área sob a curva ROC. Ela indica a capacidade do modelo em distinguir entre as classes.

V. Analise Exploratória de dados

A análise exploratória de dados é de suma importância para a implementação de algoritmos preditivos, consta-se com etapas primordiais para entender e compreender os dados de um DataSet antes da síntese do modelo. Essa fase garante uma maior qualidade e confiabilidade para o modelo assim sintetizado.

O cerne dessa análise consiste em uma clara e minuciosa investigação da distribuição dos dados, frequentemente auxiliada por meio de representações gráficas, com o intuito de compreender como cada parâmetro pode influenciar a variável de interesse. Dada a natureza diversificada dos conjuntos de dados, diversas abordagens podem ser adotadas na análise exploratória. Em muitos casos, opta-se por uma análise univariada, a qual examina individualmente cada variável. Tal análise nos permite extrair informações cruciais, tais como a media, a mediana, o desvio padrão, os valores minimo e máximo, bem como a frequência dos valores associados a uma variavel especiifica. Esses dados estatisticos revelam-se essenciais para obtermos uma compreensão profunda dos dados em questão.

A. Tamanho dos Artigos

Na Figura 1, apresenta-se a distribuição do tamanho dos artigos em número de palavras. Observa-se que a maioria dos artigos possui relativamente poucas palavras, concentrando-se, em sua maioria, abaixo de 300 palavras. Além disso, há uma longa cauda que se estende até a faixa de 4000–5000 palavras, indicando a presença de alguns artigos extensos.

A distribuição, portanto, apresenta um comportamento assimétrico, com forte concentração de artigos na faixa inicial (próxima de 100–200 palavras) e uma cauda longa. Esse tipo de comportamento é comum em conjuntos de dados textuais, em que há muitos textos curtos e poucos textos muito longos. Para a análise, é importante levar em conta essa heterogeneidade no tamanho dos artigos, pois ela pode impactar métodos de processamento de linguagem natural.

B. Top 20 Palavras Mais Frequentes: Notícias Reais vs. Notícias Falsas

A Figura 2 exibe as 20 palavras mais frequentes no conjunto de notícias classificadas como reais. Nota-se que

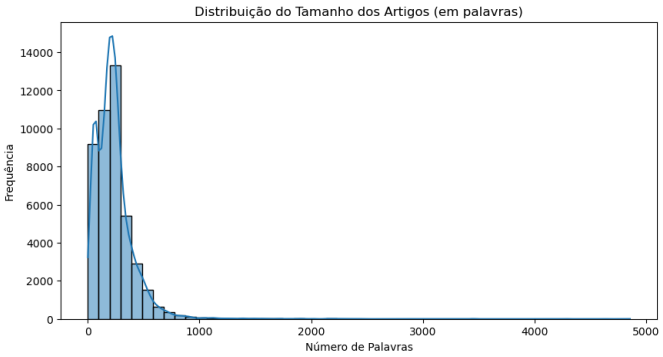

Fig. 1. Número de palavras por notícia

o termo *"said"* se destaca como o mais recorrente, indicando a presença de muitas citações diretas em matérias jornalísticas. Em seguida, aparecem termos relacionados ao contexto político norte-americano, como *"trump"*, *"president"*, *"state"* e *"government"*. O termo *"reuters"* sugere a origem ou referência a notícias de agência internacional.

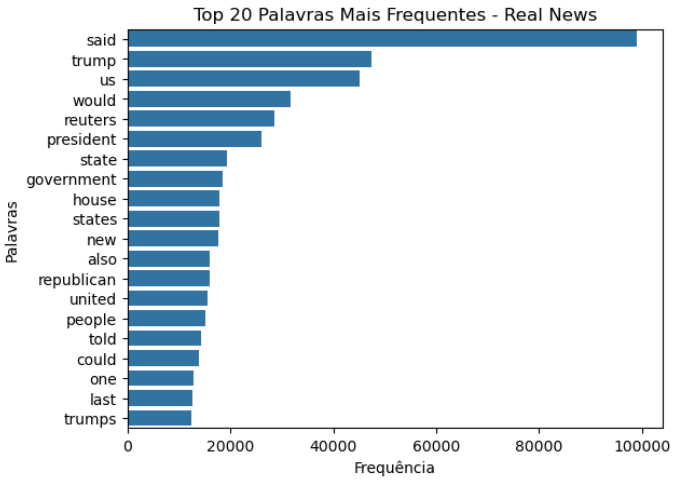

Fig. 2. Top 20 palavras mais frequentes em notícias reais.

Já a Figura 3 mostra as 20 palavras mais frequentes nas notícias classificadas como falsas. Nesse caso, *"trump"* é a palavra mais recorrente, seguida de *"said"*, *"president"* e *"people"*. Além de termos semelhantes aos da Figura 2 (*"would"*, *"us"*, *"new"* etc.), observa-se a presença de nomes específicos de figuras políticas como *"obama"*, *"clinton"*, *"donald"* e *"hillary"*, sugerindo um foco maior em personalidades políticas e possivelmente em conteúdo sensacionalista ou polémico.

1) Análise Comparativa: Comparando as duas distribuições (Figuras 2 e 3), percebe-se que:

- **Convergências:** Palavras como *"trump"*, *"said"* e *"president"* aparecem em ambos os conjuntos, indicando um interesse comum.
- **Diferenças:** Em notícias reais, termos como *"reuters"*, *"government"* e *"house"* são mais proeminentes, en-

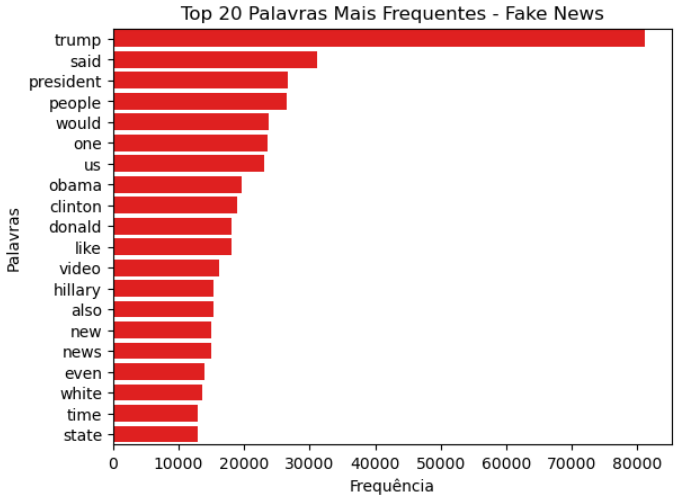

Fig. 3. Top 20 palavras mais frequentes em notícias falsas.

quanto nas notícias falsas surgem nomes específicos (*"obama"*, *"clinton"*, *"donald"*, *"hillary"*), sugerindo que o corpus de *fake news* enfatiza mais personalidades políticas e eventuais teorias ou histórias sensacionalistas.

- **Vocabulário:** Em notícias falsas, a presença de palavras como *"video"*, *"like"* e *"news"* pode indicar um formato mais informal ou opinativo. Em contrapartida, a forte incidência de *"said"* e *"reuters"* em notícias reais sugere conteúdo mais jornalístico, com uso de citações de fontes oficiais.

De forma geral, a análise comparativa das palavras mais frequentes reforça a diferença de estilo e conteúdo entre notícias reais e falsas, evidenciando que, embora haja temas em comum (como política e figuras públicas), as notícias falsas tendem a enfatizar mais nomes específicos e um tom possivelmente mais sensacionalista, enquanto as notícias reais apresentam um tom mais factual e formal.

C. Frequencia Relativa das Palavras

A frequência relativa de palavras é uma métrica importante para classificação de textos, pois ajuda a identificar padrões característicos em diferentes tipos de textos, dessa forma conseguimos capturar padrões e estruturar dados textuais de forma que algoritmos possam classificá-los de maneira eficiente.

Como podemos ver na figura 4, temos palavras que são mais utilizadas em Fake News. Por exemplo, "video" e "like", são mais comuns em Fake News, sugerindo que esses textos podem ser direcionados para engajamento e viralização em redes sociais, ao invés de informação objetiva. A alta frequência de nomes próprios de políticos em Fake News podem indicar um foco maior em narrativas polarizadas e ataques pessoais. Também temos algumas palavras que são mais frequentes em Real News, como "said" que sugere que real news utilizam mais citações diretas e entrevistas, tornando a informação mais verificável. A palavra "state" também aparece mais em Real

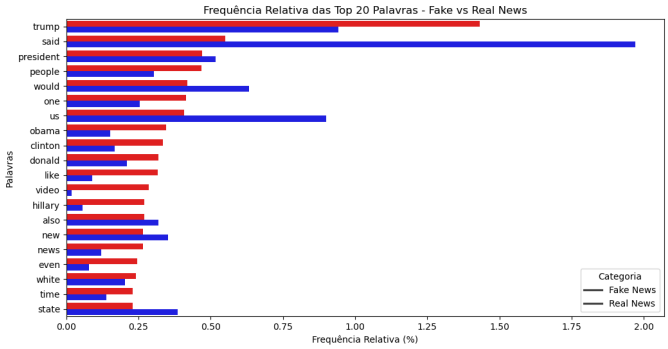

Fig. 4. Frequência Relativa das Top 20 Palavras - Fake vs Real News

News, significando menções a governo, políticas públicas e eventos institucionais.

D. Tamanho das Noticias separado por veracidade

Fake News tendem a serem curtas e diretas, projetadas para viralizar rapidamente, visto que utilizam manchetes chamativas e mensagens curtas para prender a atenção. Textos menores são mais fáceis de consumir e compartilhar em redes sociais, além de que Fake News frequentemente carecem de detalhes, fontes verificáveis e contexto, resultando em textos mais curtos.

Real News costumam ser mais extensos, por conter mais detalhes, explicações e fontes, resultando em textos mais extensos. O jornalismo profissional inclui depoimentos e estatísticas, apresentando múltiplos pontos de vista e contexto histórico.

Fatos que podem ser constatados pela figura 5, uma vez que textos Fake News mais frequentemente tem menos que 100 palavras, enquanto que Real News tem mais frequentemente 200 palavras.

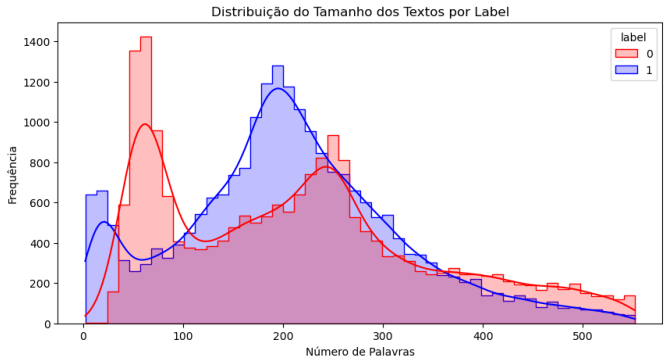

Fig. 5. Distribuição do Tamanho dos Textos por Label

VI. Resultado e Discussão

A. Multinomial Naive Bayes

O Multinomial Naive Bayes apresentou as seguintes métricas:

```
Relatório de Classificação:
              precision    recall  f1-score   support

           0       0.95      0.96      0.96      4228
           1       0.97      0.95      0.96      4752

    accuracy                           0.96      8980
   macro avg       0.96      0.96      0.96      8980
weighted avg       0.96      0.96      0.96      8980


Acurácia: 0.9589086859688196
Precision (MultinomialNB): 0.9673704414587332
AUC-ROC: 0.9836283642161127
AUC-PR: 0.9860810878125684
```

Fig. 6. Relatório de Classificação - Multinomial Naive Bayes

B. Complement Naive Bayes

Já o Complement Naive Bayes apresentou o seguinte resultado:

```
Relatório de Classificação:
              precision    recall  f1-score   support

           0       0.95      0.96      0.96      4228
           1       0.97      0.95      0.96      4752

    accuracy                           0.96      8980
   macro avg       0.96      0.96      0.96      8980
weighted avg       0.96      0.96      0.96      8980


Acurácia: 0.9586859688195991
Precision (ComplementNB): 0.9673565180285897
AUC-ROC: 0.9836275180852996
AUC-PR: 0.9860801940006345
```

Fig. 7. Relatório de Classificação - Complement Naive Bayes

Ambos os modelos se mostraram altamente eficazes na classificação de Fake News, apresentando valores muito parecidos para todas as métricas. A acurácia está em torno de 95.9% indicando que os modelos conseguem classificar as notícias com alta precisão. A precisão indica que, quando o modelo prevê que uma notícia é falsa, ele está certo em aproximadamente 96.7% das vezes. A AUC-ROC, aproximadamente 98%, mostra que os modelos possuem uma excelente separação entre classes, o que significa que conseguem distinguir bem entre Fake News e Real News.

Dado que os dois modelos têm desempenho quase idêntico, a escolha pode depender de outros fatores, como eficiência computacional, o MultinomialNB é mais simples e pode ser ligeiramente mais rápido, ou robustez a dados desbalanceados, o ComplementNB é projetado para lidar melhor com dados desbalanceados.

VII. Conclusão

Neste estudo, utilizamos dois classificadores baseados no Naive Bayes – MultinomialNB e ComplementNB – para a detecção de Fake News. O Naive Bayes é um modelo probabilístico eficiente para classificação de textos, sendo amplamente utilizado em problemas como filtragem de spam, análise de sentimentos e detecção de notícias falsas.

Os resultados mostram que ambos os modelos se mostraram altamente eficazes na classificação de Fake News, com desempenho quase idêntico em todas as métricas avaliadas.

Os resultados indicam que o Naive Bayes continua sendo uma excelente abordagem para classificação automática de Fake News, devido à sua simplicidade, eficiência e alto desempenho. Para aprimorar ainda mais os resultados, podem ser exploradas técnicas como engenharia de características, TF-IDF para vetorização de texto, ou até mesmo a combinação com outros classificadores mais complexos.

VIII. Referências

[1] What is exploratory data analysis (EDA)?https://www.ibm.com/topics/exploratory-data-analysis

[2] K. R. et al. *Naive Bayes Classifiers for Fake News Detection*. IEEE Access, 2020.

[3] Ahmed H, Traore I, Saad S. *Detecting opinion spams and fake news using text classification*. Journal of Cybersecurity, 2018.

[4] Loper, E., & Bird, S. *NLTK: The Natural Language Toolkit*. arXiv preprint cs/0205028, 2002.

[5] Pedregosa et al. *Scikit-learn: Machine Learning in Python*. JMLR 12, pp. 2825-2830, 2011.

[6] Kaggle. Fake and Real News Dataset. https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset