Classificação de Fake News Utilizando o Dataset WELFake

Marlon Silva Pereira
Insper - Instituto de Ensino e Pesquisa, São Paulo, Brasil
Email: marlonsp@al.insper.edu.br

***Abstract*—A proliferação de notícias falsas nas redes sociais tem sido um problema crescente, uma vez que as informações são frequentemente compartilhadas sem a verificação de sua autenticidade. Este projeto visa analisar a detecção de notícias falsas utilizando o dataset WELFake, com base nos métodos e resultados apresentados no artigo "WELFake: Word Embedding Over Linguistic Features for Fake News Detection". O artigo propõe um modelo de classificação em duas fases que utiliza embeddings de palavras e características linguísticas para identificar a veracidade de notícias, atingindo uma precisão de 96,73%. A análise feita neste projeto se baseia nos resultados apresentados nesse estudo, além de utilizar classificadores como Naive Bayes, SVM e Random Forest. A escolha do SVM se deve aos resultados superiores obtidos no paper de referência, servindo como base para comparações.**

I. Dataset

O dataset WELFake, publicado em IEEE Transactions on Computational Social Systems, contém 78.098 entradas divididas em quatro colunas: número de série, título, texto e rótulo (0 = fake, 1 = real). Ele serve como base para a detecção de fake news, contendo exemplos reais e falsos de notícias. O pré-processamento incluiu a substituição de valores ausentes e a combinação de título e texto para formar a variável *combined_text*.

II. Pipeline de Classificação

A. Pré-processamento

O texto foi submetido a uma série de transformações, incluindo:

- **Lematização**: Redução de palavras à sua forma base para uniformizar o vocabulário.
- **Stemming**: Remoção de sufixos para reduzir as palavras à sua raiz.
- **Remoção de stopwords**: Palavras sem valor semântico para a tarefa de classificação foram excluídas.
- **Tokenização**: Separação do texto em tokens (palavras individuais).

Após o pré-processamento, foi utilizado a técnica de TF-IDF para transformar o texto em vetores numéricos, representando a relevância das palavras em cada documento.

B. Modelo de Classificação

Optou-se pelo uso do SVM (*Support Vector Machine*) com kernel linear e ponderação para classes desbalanceadas. Esse modelo foi escolhido por seu desempenho em problemas de classificação binária, especialmente com dados textuais.

III. Avaliação das Palavras mais importantes

As palavras que mais influenciaram a decisão do classificador foram extraídas a partir dos coeficientes do modelo SVM. Foram observadas as cinco palavras com maior impacto positivo e negativo na classificação.

A. Top 5 palavras com maior coeficiente positivo:

- via: 15.5107
- imag: 9.6154
- video: 5.4752
- octob: 5.4620
- hillari: 4.6484

B. Top 5 palavras com menor coeficiente negativo:

- reuter: -22.4815
- 000: -10.9173
- breitbart: -10.2780
- said: -8.9144
- twitter: -6.7402

A análise dos coeficientes do modelo SVM revelou que palavras relacionadas a mídias visuais, como "via", "imag" e "video", têm grande impacto positivo na identificação de notícias falsas. Já palavras ligadas a fontes de notícias, como "reuter" e "breitbart", apresentam coeficientes negativos significativos, sugerindo que menções a fontes estabelecidas estão mais associadas à classificação de notícias verdadeiras. Isso indica que o classificador tende a se apoiar em termos específicos para diferenciar entre conteúdos reais e falsos.

IV. Avaliação do Classificador

A métrica de avaliação utilizada foi o Balanced Accuracy, para compensar o desbalanceamento das classes. A avaliação inicial do modelo resultou em uma acurácia balanceada de 95.83%. Isso mostra um resultado muito próximo do esperado, com relação ao trabalho de referência, no qual foi obtido uma acurácia de 96.73% com o mesmo modelo.

V. Tamanho do Dataset

Foram realizadas múltiplas execuções variando o tamanho da amostra de treinamento (10%, 25%, 50%, 75% e 100% do dataset) para entender o impacto do tamanho do conjunto de dados na acurácia do modelo. O gráfico a seguir ilustra a relação entre o tamanho do dataset e a acurácia obtida.

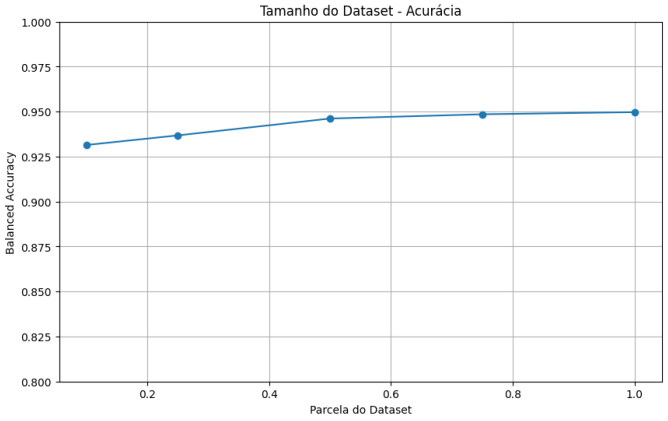

Fig. 1. Influência do tamanho do Dataset no resultado da acurácia obtida, indicando um aumento no ínicio da variação do tamanho, mas uma estabilização da acurácia com uma porcentagem maior de amostra do Dataset

Além disso, foi comparado também o erro de treino e teste para diferentes tamanhos de dataset, observando a tendência de overfitting com conjuntos de treino pequenos.

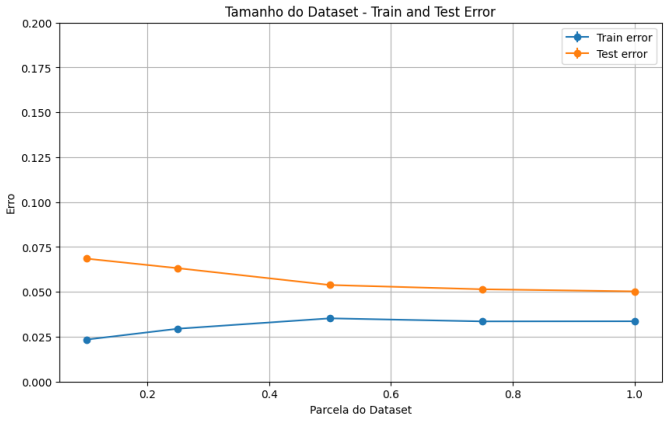

Fig. 2. Evolução do erro de treino e teste com o aumento do tamanho do dataset, mostrando uma redução inicial no erro de teste, seguida de estabilização

Os resultados mostraram que, embora o aumento do tamanho do dataset leve a uma melhora na acurácia do modelo, essa melhora tende a se estabilizar em amostras maiores. Para o caso de uso proposto, o tamanho atual do dataset é suficiente para fornecer resultados precisos, e a expansão do dataset pode não trazer ganhos significativos além de um certo ponto.

VI. Análise de Tópicos

Foi utilizado o modelo de *Latent Dirichlet Allocation* (LDA) para identificar tópicos principais no dataset e avaliar se a acurácia de classificação varia entre esses tópicos. Foram definidos cinco tópicos e foi observado que alguns tópicos apresentaram melhor desempenho de classificação do que outros.

- **Tópico 0**: Discussões sobre relações internacionais e diplomacia, com foco em China, Estados Unidos e questões militares. Acurácia = 0.9590
- **Tópico 1**: Questões legais e judiciais envolvendo figuras políticas como Trump, além de segurança e aplicação da lei. Acurácia = 0.9481
- **Tópico 2**: Reflexões sobre eventos de longo prazo, como anos, vida, e questões relacionadas a mudanças ao longo do tempo. Acurácia = 0.9346
- **Tópico 3**: Política americana, especialmente eleições e partidos políticos, com ênfase em Trump e Clinton. Acurácia = 0.9339
- **Tópico 4**: Cobertura da mídia sobre Donald Trump e Hillary Clinton, com foco em percepções populares e mídias sociais. Acurácia = 0.9565

Os resultados indicam que alguns tópicos estão mais relacionados a padrões de fake news, enquanto outros são mais desafiadores para o classificador. Tópicos como relações internacionais (0.9590) e cobertura midiática de Trump e Clinton (0.9565) tiveram melhor desempenho, enquanto temas mais amplos, como política americana (0.9339) e eventos de longo prazo (0.9346), foram mais desafiadores para o classificador, indicando que a precisão varia de acordo com o conteúdo abordado.

VII. Conclusão

Este estudo replicou os resultados de classificação de fake news utilizando o dataset WELFake, com foco no modelo SVM, alcançando uma acurácia balanceada de 95.83%. A análise de tópicos mostrou que certos temas são mais facilmente classificados como fake ou real, sugerindo que o conteúdo abordado tem impacto na precisão da classificação.

References

[1] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "WELFake: Word Embedding Over Linguistic Features for Fake News Detection," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 881–893, 2021. doi:10.1109/TCSS.2021.3068519.