Impacto de Representações Textuais em Modelos de Classificação de *Fake News*: Uma Abordagem Comparativa

Ada Maris Pereira Mário

[1]Relatório desenvolvido para a
Disciplina de Mineração de Dados não Estruturados - ICMC-USP

Resumo. *Este projeto explora a eficácia de diferentes técnicas de pré-processamento e representação de textos na tarefa de classificação de notícias falsas. Foram comparados os métodos TF-IDF, Word2Vec e duas variantes do BERT (Sentence-BERT e DistilBERT), aplicados ao WELFake, um dataset de notícias reais e falsas. Utilizando o algoritmo k-Nearest Neighbors, foram analisados os desempenhos para cada representação, avaliados por meio do teste de Wilcoxon para verificar diferenças significativas entre as técnicas. Os resultados indicam que a versão SBERT apresenta um desempenho não diferente das técnicas tradicionais, enquanto que a versão mais compacta DistilBERT se sobressai em meio aos demais.*

1. Identificação do Problema

Notícias falsas são definidas como informações fabricadas que imitam o conteúdo da mídia de notícias na forma, mas não no processo organizacional ou intenção. Seus veículos, por sua vez, não têm as normas e processos editoriais da mídia de notícias para garantir a precisão e a credibilidade das informações [Lazer et al. 2018]. Apesar de *fake news* não serem um fenômeno novo, o advento das mídias sociais como parte integrada às vidas das pessoas proporcionou um ambiente fértil para o compartilhamento rápido de informações, de modo a tornar difícil a diferenciação entre fatos confirmados e informações de baixa qualidade com dados propositalmente falsos [Shu et al. 2017].

As notícias falsas atraíram principalmente a atenção recente em um contexto político, mas também foram documentadas em informações promulgadas sobre tópicos como vacinação, nutrição e valores de ações. Elas são particularmente perniciosas porque são parasitárias dos veículos de notícias padrão, beneficiando-se e minando simultaneamente sua credibilidade [Lazer et al. 2018]. Nesse sentido, um crescente número de pesquisas têm concentrado seus esforços na identificação de informações dúbias em plataformas online com o desenvolvimento de técnicas efetivas e automáticas de detecção de *fake news* por meio de inteligência artificial [Ouassil et al. 2022].

Entretanto, abordagens tradicionais de mineração de dados são inefetivas em lidar com tarefas como classificação de textos, dada a necessidade de estruturá-los. Assim, têm-se técnicas de representações de tais dados que se provaram notavelmente bem-sucedidas sem ser preciso o entendimento de propriedades específicas, como gramática e significado das palavras [Weiss et al. 2010]. Um dos principais temas que sustentam a mineração de textos é a transformação de tais em dados numéricos, portanto, embora a apresentação inicial seja diferente, os dados passam para uma representação clássica de mineração de dados.

Dessa forma, este projeto visa avaliar se para diferentes pré-processamentos dos textos — aqui serão usados os modelos espaço-vetoriais TF-IDF e *Word2Vec*, bem como modelos de linguagens neurais da família BERT [Devlin 2018] — há diferenças significativas no desempenho de tarefas de classificação que tomam como base para a estruturação do modelo a própria representação do conjunto de dados, neste caso o algoritmo *k-Nearest Neighbors*. Para avaliar as hipóteses (1), será utilizado o teste de Wilcoxon para dados pareadas ao nível $\alpha = 5\%$ de significância, de modo a testar par a par, unilateralmente à direita, as métricas obtidas com cada representação. Espera-se como resultado um desempenho dos modelos de pelo menos 80% de acurácia no conjunto de teste, mas com os métodos de pré-processamento mais sofisticados obtendo um melhor desempenho que os métodos clássicos (ou seja, espera-se rejeitar a hipótese nula).

$$
\begin{aligned}
&E(D) : \text{Esperança da diferença entre as métricas do } 2^\circ \text{ menos do } 1^\circ \text{ modelo} \\
&H_0 : E(D) \leq 0 \quad \text{(O desempenho do } 2^\circ \text{ modelo é igual ou pior que o } 1^\circ\text{)} \\
&H_1 : E(D) > 0 \quad \text{(Há diferenças nos desempenhos dos modelos — o } 2^\circ \text{ modelo é melhor)}
\end{aligned}
\tag{1}
$$

O conjunto de dados para o desenvolvimento deste projeto corresponde ao WELFake Dataset [Verma et al. 2021], com mais de 72 mil artigos de notícias em lingua inglesa, sendo mais de 37 mil falsas e mais de 35 mil reais. O dataset é resultado de uma mescla de diversos datasets populares de notícias, com a maioria relacionada a política, em especial política norte-americana.

2. Pré-processamento

2.1. Amostragem e Tokenização

Os dados aqui utilizados são organizados com três atributos: `'title'`, com valores correspondentes aos títulos das notícias; `'text'`, com os textos dos corpos das notícias; e `'label'` com valores binários apontando as classes, com 0 indicando notícias falsas e o valor 1 indicando notícias reais.

Para melhor manuseamento e processamento dos textos para vetores, retirou-se uma amostra de 10 mil entradas do conjunto total, descartando-se possíveis valores nulos ou inválidos, bem como mantendo o equilíbrio de classes, de modo a se ter 5 mil entradas válidas em cada uma. Em seguida, dividiu-se o conjunto em treino e teste, com 2/3 dos separados para a etapa de treinamento e extração de padrões pelos modelos, e o restante para predição. Ademais, mesclaram-se as colunas `'title'` e `'text'` em uma denominada `'combined'`, que por sua vez foi processada para tokenização, criando a coluna `'processed'` com vetores de strings para uso dos métodos espaço-vetoriais.

A partir de tais vetores pôde-se fazer uma análise exploratória das notícias. Na Figura 1 vê-se as palavras mais frequentes da amostra ordenadas para cada classe. Como mencionado anteriormente, a maior parte das notícias é relacionada à política estadunidense, como é refletido na alta frequência de termos ou figuras específicas dessa temática em ambas as classes. No entanto, observa-se que as notícias falsas apresentam frequências absolutas maiores de certas palavras, como *"said"* e *"trump"*, podendo-se supor uma recorrência de tema ou tentativa de reforço de uma narrativa específica.

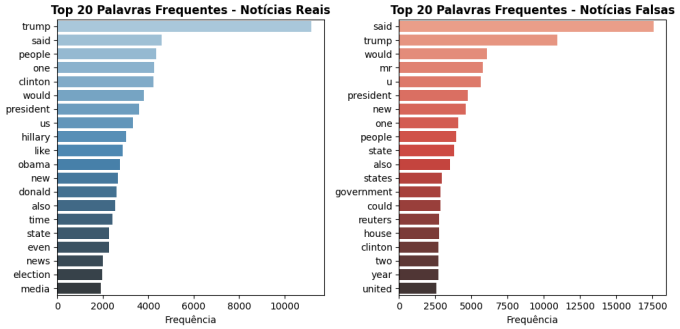

Figura 1. 20 Palavras mais frequentes segmentadas por classes

2.2. TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) é uma medida estatística que representa quão influente uma palavra é para definir a "relevância" de um documento em um corpus [Ramos et al. 2003]. Dada uma coleção de documentos D, uma palavra w e um documento individual $d \in D$, calcula-se o valor definido por (2), onde $f_{w,d}$ é igual ao número de vezes que w aparece em d, $|D|$ é o tamanho do corpus e $f_{w,D}$ é igual ao número de documentos nos quais w aparece em D. Assim, seu valor aumenta proporcionalmente à frequência com que uma palavra aparece em um documento e é compensado pelo número de documentos no corpus que também contêm essa palavra.

$$w_d = f_{w,d} * log(|D|/f_{w,D}) \tag{2}$$

Para a ponderação dos dados na coluna `'processed'`, consideraram-se apenas trigramas, de forma a capturar contextos mais específicos nas frases, em vez de palavras isoladas ou bigramas, que podem ser mais ambíguos. Limitou-se o vocabulário a 5000 principais termos, ordenados pelo valor de TF-IDF. Desse modo, a representação final dos dados consistiu em uma matriz esparsa de dimensão $N \times 5000$, onde N representa o número de documentos no conjunto de dados (6,7 mil no conjunto de treino e 3,3 mil no conjunto de teste). Cada linha dessa matriz corresponde a um documento, e cada coluna corresponde ao valor TF-IDF de um trigrama específico, indicando a relevância de termos para a classificação.

2.3. *Word2vec*

O *Word2Vec* é um modelo de representação de palavras em um espaço vetorial contínuo, onde a proximidade entre vetores reflete a semelhança semântica das palavras [Mikolov 2013]. Neste projeto, foi utilizado o modelo pré-treinado pelo corpus do *Google News*, com 3 milhões de vetores de palavras em inglês de 300 dimensões.

Para representar cada documento usando *Word2Vec*, é necessário combinar os vetores das palavras, criando embeddings de documentos que capturam informações agregadas de cada sentença ou parágrafo. Aqui, foram utilizados embeddings médios, ou seja, cada notícia foi representada pela média dos vetores das palavras tokenizadas contidas no texto. Essa abordagem resulta em um vetor denso fixo em que cada documento reflete uma combinação de todas as palavras presentes na notícia, capturando uma noção do conteúdo semântico médio.

Assim, cada documento de `'processed'` foi iterado e, para cada palavra que possuía um embedding pré-treinado no modelo, seu vetor foi extraído e adicionado à lista de vetores do documento. Em seguida, calculou-se a média desses vetores para cada documento, gerando um único vetor de 300 dimensões que representa cada notícia. A representação final dos dados, portanto, é uma matriz densa de dimensão $N \times 300$, onde N representa o número de documentos em cada conjunto.

2.4. Sentence-BERT

O *Sentence-BERT* (SBERT) é uma modificação do BERT (*Bidirectional Encoder Representations from Transformers*) especialmente projetada para produzir embeddings de sentenças e documentos inteiros, ao contrário do BERT original, que gera embeddings para cada palavra individualmente e requer um pós-processamento adicional para obter uma representação única para textos maiores [Reimers 2019]. Diferentemente dos métodos clássicos, que se baseiam principalmente nas frequências e coocorrências de palavras, os modelos BERT capturam relações semânticas mais complexas entre as palavras, embasadas em contexto de frase e significado global.

O modelo usado neste projeto, `paraphrase-MiniLM-L6-v2`, é uma versão compacta e otimizada do SBERT que gera embeddings de 384 dimensões. Para cada documento no corpus, representado na coluna `'combined'`, foi gerado um embedding vetorial único de 384 dimensões, capturando as relações semânticas das frases em cada notícia.

2.5. DistilBERT

O *DistilBERT* é uma versão otimizada e mais leve do modelo BERT. Usando uma técnica chamada distilação de conhecimento, o DistilBERT mantém apenas 6 das 12 camadas do BERT original, reduzindo o número de parâmetros em aproximadamente 40%, o que resulta em ganhos consideráveis de velocidade e menor uso de memória, enquanto preserva 97% da capacidade de compreensão semântica do modelo BERT base [Sanh 2019].

É particularmente adequado para tarefas de classificação, como identificação de sentimentos e categorização de textos, pois a distilação ajuda a priorizar as representações mais relevantes dos tokens para essas tarefas. Diferente do SBERT, que adapta o BERT para produzir embeddings de sentenças e é mais indicado para tarefas de similaridade semântica, o DistilBERT retém a estrutura tokenizada do BERT. Essa estrutura oferece uma representação mais detalhada no nível de tokens, especialmente vantajosa quando se depende de padrões locais dentro de um texto, aumentando a acurácia em tarefas de detecção e categorização.

Para representar os documentos no corpus, o DistilBERT foi configurado para gerar embeddings com base no primeiro token especial [CLS] de cada sentença, que, no BERT e DistilBERT, serve como uma agregação da informação contida em toda a sequência. A representação final dos documentos com DistilBERT resultou em uma matriz de dimensão $N \times 768$, onde N é o número de documentos no corpus.

3. Extração de Padrões

Nesta seção, são detalhadas as técnicas aplicadas para extrair padrões dos da-

dos com base nas representações vetoriais das notícias. A princípio, foram realizadas projeções das *embeddings* com *Uniform Manifold Approximation and Projection* (UMAP). O UMAP é uma técnica de redução de dimensionalidade que preserva as relações de vizinhança entre os dados ao projetá-los em um espaço de menor dimensão [McInnes et al. 2018]. Ele funciona otimizando uma estrutura de grafo que representa as similaridades locais dos dados, preservando tanto a estrutura global quanto os agrupamentos locais. No caso deste projeto, o UMAP foi utilizado para projetar as *embeddings* de cada técnica em um espaço bidimensional (2D), facilitando a visualização de padrões e proximidades semânticas entre documentos.

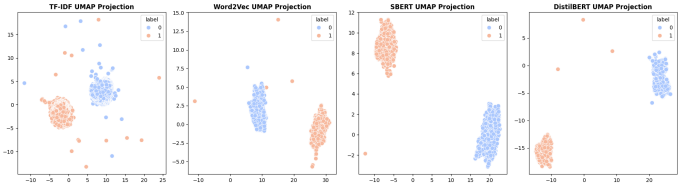

Figura 2. Projeções UMAP 2D para cada representação

Na Figura 2 pode-se observar que em geral as distribuições dos documentos por classe pelas representações estão bem delineadas, com as técnicas mais clássicas apresentando maior proximidade e algumas sobreposições entre os pontos de diferentes categorias.

Para a tarefa de classificação das notícias com base no pré-processamento, usou-se o k-Nearest Neighbors, uma vez que tal algoritmo não requer suposições de distribuição prévia dos dados, o que o torna particularmente adequado para comparar *embeddings* de texto, onde a proximidade entre vetores reflete a similaridade semântica. Para obter melhor desempenho, realizou-se o ajuste de hiperparâmetros, identificando-se a combinação ideal de número de vizinhos (k) no intervalo $[1, 20]$ e métrica dentre as opções de distâncias euclidiana, Minkowski, Manhattan e similaridade de cossenos. Na Tabela 1 apresentam-se os valores encontrados para cada modelo.

Tabela 1. Melhores parâmetros para os modelos

Representação	Métrica	k	*Score*
TF-IDF	Cosseno	17	$0,86$
Word2Vec	Cosseno	19	$1,0$
SBERT	Cosseno	17	$1,0$
DistilBERT	Manhattan	10	$1,0$

4. Pós-processamento

Para avaliar o desempenho dos modelos utilizados neste projeto, foram calculadas as métricas de acurácia (proporção de previsões corretas), precisão (nível de exatidão entre as previsões positivas), recall (capacidade do modelo de identificar corretamente as instâncias positivas), F1-score (média harmônica entre precisão e recall, ponderando o equilíbrio entre os dois), e AUC-ROC (área sob a curva ROC, que mede a capacidade do

modelo em distinguir entre classes). Além disso, foi realizada uma validação cruzada em cinco subdivisões dos dados para cada técnica, gerando uma média de desempenho para cada representação vetorial. Na Tabela 2, são apresentados tais valores.

Tabela 2. Métricas de Desempenho dos Modelos

Representação	Acurácia	Precisão	Recall	F1-Score	AUC-ROC	CV
TF-IDF	$0,8394$	$0,8303$	$0,8457$	$0,8379$	$0,9067$	$0,7964$
Word2Vec	$0,8403$	$0,8700$	$0,7932$	$0,8298$	$0,9187$	$0,8127$
SBERT	$0,8294$	$0,8430$	$0,8019$	$0,8219$	$0,9071$	$0,8036$
DistilBERT	$0,8952$	$0,9158$	$0,8660$	$0,8902$	$0,9590$	$0,8833$

A análise inicial revela que o modelo com DistilBERT obteve o melhor desempenho geral, especialmente em validação cruzada e na métrica AUC, o que indica uma alta capacidade de discriminar entre as classes de forma mais robusta. Por outro lado, o modelo com SBERT teve algumas métricas piores ou muito similares com os valores dos modelos com TF-IDF e *Word2vec*. Para avaliar se de fato essas diferenças são significativas, foi realizado o teste de Wilcoxon com base nas hipóteses definidas em 1 ao nível $\alpha = 0,05$ de significância. Tal teste estatístico não-paramétrico é ideal para comparações de pares de medidas de uma mesma amostra, ou seja, quando seus elementos são medidos em duas ocasiões ou sob duas condições diferentes — aqui as condições sendo as diferentes representações vetoriais. A Tabela 3 apresenta os níveis descritivos obtidos com os testes entre as combinações de modelos.

Tabela 3. Resultados do teste de Wilcoxon entre as representações

Diferença entre métricas - D	**Valor-*p*** - $Pr(D \geq d)$
Word2Vec − TF-IDF	$0,5$
SBERT − TF-IDF	$0,84375$
DistilBERT − TF-IDF	$0,03125$
SBERT − *Word2Vec*	$0,9375$
DistilBERT − *Word2Vec*	$0,03125$
DistilBERT − SBERT	$0,03125$

Como mostrado, focalizou-se em testar os métodos mais sofisticados e otimizados contra as opções clássicas ou menos aprimoradas. Os resultados mostram que as comparações envolvendo DistilBERT resultaram em valores-$p \leq 0,05$, rejeitando a hipótese nula, o que indica uma diferença estatisticamente significativa em relação aos outros métodos, confirmando seu superior desempenho. Em contrapartida, todas as comparações com SBERT tiveram valores-*p* muito altos, indicando que de fato o modelo teve desempenho igual ou pior que os modelos com métodos tradicionais, uma discordância com um dos resultados esperados definidos no princípio. Possíveis razões para isso é que o SBERT foi otimizado para tarefas de emparelhamento de sentenças e recuperação de informações, que podem não estar diretamente alinhadas com a tarefa de classificação de notícias em fake e real deste projeto. Além disso, o SBERT pode precisar de ajustes específicos, como *fine-tuning* com dados similares para obter melhor desempenho em tarefas de classificação.

De modo geral, os modelos obtiveram desempenhos iguais ou superiores ao objetivo proposto no início. Para refinar tais resultados, pode-se ajustar os parâmetros para as configurações específicas dos dados e explorar outras métricas de análise mais avançadas ou específicas, visando superar os objetivos estabelecidos.

5. Uso do Conhecimento

O conhecimento extraído neste projeto oferece suporte estratégico para identificar e combater a desinformação em diversos contextos. Com um desempenho elevado, o modelo DistilBERT demonstrou alta capacidade de distinguir entre classes, sendo ideal para sistemas que priorizam precisão na triagem de conteúdos suspeitos, especialmente em cenários de alto volume informacional, a exemplo de períodos eleitorais. Em situações em que há restrições computacionais, modelos mais simples como TF-IDF e Word2Vec, que apresentaram desempenho razoável, mostraram-se igualmente viáveis para soluções leves.

Para aprimoramentos futuros, é essencial a investigação dos casos de erro recorrentes em cada técnica de representação vetorial. Analisar as *fake news* que escaparam de todos os modelos ajudaria a entender limitações na detecção e a inspirar novos métodos de representação e construção de classificadores, de modo a contribuir para o refinamento contínuo dos modelos e impulsionar o desenvolvimento de soluções que atendam melhor as necessidades de diferentes contextos.

Referências

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mikolov, T. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Ouassil, M.-A., Cherradi, B., Hamida, S., Errami, M., EL GANNOUR, O., and Raihani, A. (2022). A fake news detection system based on combination of word embedded techniques and hybrid deep learning model. *International Journal of Advanced Computer Science and Applications*, 13(10).

Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

Reimers, N. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Sanh, V. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Verma, P. K., Agrawal, P., Amorim, I., and Prodan, R. (2021). Welfake: word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, 8(4):881–893.

Weiss, S. M., Indurkhya, N., Zhang, T., and Damerau, F. (2010). *Text mining: predictive methods for analyzing unstructured information*. Springer Science & Business Media.