Avaliação de Desempenho Das Etapas do Processo de Extração de Conhecimento para Detecção de *Fake News* em Português

Alice Barbosa, Felipe Sousa, Reinaldo Braga

[1]Laboratório de Redes de Computadores e Sistemas (LAR)
Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)

Abstract. *Due to the significant increase in fake news on social networks, several studies and strategies have been developed in recent years in order to identify them. Therefore, this article presents a proposal for the automatic identification of fake news. More specifically, an analysis of the algorithm construction process is carried out, investigating the impact of decision-making in the methodological process on the final result. For this purpose, the Fake.Br database was used, which presents 7,200 news articles in Portuguese. It is noteworthy that the research carried out focused on analyzing both the texts and their respective metadata. Thus, after analyzing the combinations, an average accuracy of 97% was obtained.*

Resumo. *Devido ao aumento significativo de fake news nas redes sociais, diversos estudos e estratégias têm sido desenvolvidos nos últimos anos com o intuito de identificá-las. Sendo assim, o presente artigo apresenta uma proposta para identificação automática de fake news. Mais especificamente, é realizada uma análise do processo de construção dos algoritmos, investigando o impacto das tomadas de decisão no processo metodológico no resultado final. Para este fim, foi utilizada a base de dados Fake.Br, que apresenta 7.200 artigos de notícias em português. Destaca-se que a pesquisa realizada concentrou-se em analisar tanto os textos, quanto os seus respectivos metadados. Desta forma, após uma análise das combinações, obteve-se uma média da acurácia de 97%.*

1. Introdução

O avanço da tecnologia tem trazido novas práticas na vida social, proporcionando meios modernos de comunicação, oferecendo facilidade e celeridade na realização de tarefas cotidianas. Esta realidade representa um fator importante para o aumento do número de usuários ativos que acessam a Internet regularmente no mundo. Segundo uma pesquisa publicada pela [DataReportal 2022], atualmente, existem 5,16 bilhões de usuários de Internet no mundo, o que significa que 64,4% da população total do mundo está online.

Em paralelo a este cenário, percebe-se também a evolução das redes sociais. O estudo da [DataReportal 2022] revela ainda que, existem 4,76 bilhões de usuários de mídias sociais em todo o mundo, o que representa quase 60% da população global total. Destaca-se que o crescimento de usuários de mídia social desacelerou nos últimos tempos, com a adição líquida de 137 milhões de novos usuários neste ano, o que equivale a um crescimento anual de apenas 3%.

Ao mesmo tempo que o avanço das redes sociais tem sido importante para a sociedade, elas também têm impulsionado o consumo de notícias *online*. Em uma pesquisa

realizada pelo [Reuters and of Oxford 2021] em 12 países, cerca de 66% dos entrevistados declararam fazer uso de duas ou mais redes sociais para consumir, compartilhar ou discutir notícias. Entre as mais citadas estão WhatsApp, Instagram, TikTok, Telegram e Facebook, respectivamente.

Entretanto, a facilidade de acesso, a velocidade com que as informações são compartilhadas, aliadas ao baixo custo, também proporcionam um espaço propício para a proliferação de notícias falsas. Segundo [Fuller et al. 2009], com o crescimento maciço da comunicação *online*, o potencial para enganar as pessoas por meio da Comunicação Mediada por Computador (CMC) também cresceu, e esse engano pode resultar em danos desastrosos de longo alcance em muitas áreas do cotidiano, incluindo os mercados financeiros [Carvalho et al. 2011], [Kogan et al. 2019] e eventos políticos [Allcott and Gentzkow 2017], [Aro 2016].

Os autores [Allcott and Gentzkow 2017] definem *fake news* como artigos de notícias cujos conteúdos são intencional e comprovadamente falsos, sendo capazes de enganar os leitores, muitas vezes para a obtenção de vantagens financeiras, ou por motivações ideológicas, como o favorecimento de certos candidatos a cargos políticos em campanhas eleitorais. De acordo com os autores [Kumar et al. 2021], a difusão e a acessibilidade da Internet facilitam o acesso às informações para as grandes massas em um clique. Porém, essa acessibilidade à rede não oferece aos usuários as mesmas facilidades para a verificação da veracidade das informações.

Devido ao risco que a propagação de *fake news* representa à democracia, muitos esforços têm sido empenhados no sentido de desenvolver estratégias para a identificação de notícias falsas e, assim, minimizar os seus danos à sociedade [Sharma et al. 2019]. Neste sentido, boa parte dos estudos que buscam apresentar soluções a este problema fazem uso das técnicas de Processamento de Linguagem Natural (PLN), que segundo[Gonzalez and Lima 2003] trata-se computacionalmente dos diversos aspectos da comunicação humana, como sons, palavras, sentenças e discursos, considerando formatos e referências, estruturas e significados e contextos e usos.

Tendo em vista os impactos nocivos que as *fake news* podem vir a ocasionar, diversos estudos desenvolveram diferentes abordagens na busca de detectar notícias falsas com alta precisão. Porém, os resultados obtidos por meio da análise das métricas de validação acabam sendo semelhantes. Sendo assim, o presente artigo apresenta uma proposta para identificação automática de *fake news*. Mais especificamente, é realizado uma avaliação de desempenho das etapas do processo de extração de conhecimento para detecção de *fake news* em português, ou seja, é analisado o processo de construção dos algoritmos, investigando o impacto das tomadas de decisão no processo metodológico no resultado final, proporcionando assim, encontrar a melhor combinação na literatura atual para a detecção de notícias falsas escritas em português.

Para alcançar este objetivo foi desenvolvido uma metodologia dividida em quatro etapas, sendo estas: I) Mapeamento Bibliográfico; II) Base de dados; III) Construção dos cenários e IV) Testes e validações. Destaca-se que, para os testes, foi utilizada a mesma base de dados, intitulada Fake.Br Corpus, que apresenta 7.200 artigos de notícias escritas em português.

Este artigo traz como contribuição a análise não apenas do texto e metadados, mas

também do processo de construção dos algoritmos, identificando os pontos que causam maior impacto no resultado final. Frisa-se que também como um diferencial este estudo foca em textos escritos em português, haja vista que grande parte dos estudos nesta área são focados no Inglês.

2. Metodologia

Como mencionado, este trabalho realiza uma análise do processo de construção dos algoritmos. Para isto, foi desenvolvido uma metodologia dividida em quatro etapas, sendo estas: I) Mapeamento Bibliográfico; II) Base de dados; III) Construção dos cenários e IV) Testes e validações.

2.1. Mapeamento Bibliográfico

Para a construção dos cenários foi levado em consideração dois pontos de observação, sendo estes: I) Estrutura técnica e II) O que já há na literatura na área de detecção de *fake news*. Desta forma, a partir de um mapeamento bibliográfico foram selecionados quatro estudos e destes extraído as decisões para a construção dos algoritmos, como pode ser visualizado na Tabela 1. Os estudos foram:

O estudo publicado por [Sousa et al. 2022], que teve como objetivo apresentar uma análise de notícias em português e a detecção de *fake news*, utilizando Aprendizagem de Máquina(AM) e Redes Neurais Convolucionais(RN). Destaca-se que o estudo concentrou-se em analisar tanto os textos, como também os metadados. Dentro dos processos para a construção do algoritmos foram utilizados como pré-processamento a remoção de *stopword* e a de remoção de caracteres estranhos, em seguida, como transformador foi aplicado apenas o *Glove*. É importante frisar que o treinamento foi realizado com quatro algoritmos, sendo estes: *Convolutional Neural Network*(CNN), *Naïve Bayes*(NB), *Support Vector Machine*(SVM) e *Multilayer perceptron*(MLP). Desta forma, após uma análise sobre os algoritmos selecionados, obteve-se uma acurácia de 97%.

O trabalho [Rosa et al. 2019], por sua vez, propõe uma solução para a detecção das *fake news*, utilizando 2 técnicas de AM, no caso, CNN e as Redes Neurais Artificiais de Múltiplas Camadas (RNA).No que diz respeito ao pré-processamento, o estudo fez uso da remoção de *stopword*, assim como remoção de caracteres estranhos, de "Noise Corpus", *stemming* e a frequência baixa. Para comprovar seus resultados o estudo exibe uma análise quantitativa da acurácia, precisão, *recall*, *F1-Score*, bem como da matriz de confusão, na qual é possível observar que a CNN se sobressai a RNA, com a acurácia de 96% e 95%, respectivamente. Em contrapartida, o trabalho faz análise apenas de textos.

A pesquisa de [Almeida et al. 2021] busca apresentar uma solução para identificar as notícias falsas brasileiras em um contexto político. Para isto, os autores propõem uma investigação quanto a qual algoritmo de aprendizado de máquina, entre SVM e NB, atinge o melhor resultado. Desta forma, foi adotada a base de dados e uma própria, assim como para o pré-processamento foi aplicado a remoção de dados com conteúdo vazio, assim como de acentos e caracteres especiais, tokenização, remoção de *stop words* e a remoção de números e pontuações. É importante ressaltar que, no processo de transformação foram aplicados o *Bag of words*(BOW) e o *TF - IDF*. O melhor desempenho foi alcançado pela combinação de SVM (RBF) + BOW com 80% de acurácia.

O artigo publicado por [de Souza et al. 2020] propõe um método que visa realizar a classificação a nível de gramática e de sentimentos baseando-se na polaridade em textos

escritos em português. Para este fim os autores utilizam também a base Fake.br Corpus, como pré-processamento é aplicado a classificação gramatical, extração de símbolos irrelevantes, classificação de polaridade e a classificação de emoções. No que diz respeito ao treinamento foram testados os algoritmos NB, *Gradient Boost*(GB), *AdaBoost*(AB), SVM e *K-nearest neighbors*(KNN). Sendo que o algoritmo que apresentou o melhor desempenho foi o GB com 92,53% de acurácia.

Tabela 1. Trabalhos selecionados.

	Nível de análise	Pré-processamento	Transformação	Treinamento	ACC
SBRC 2022	Texto + Metadados	**Remoção de stopword e caracteres estranhos**	Glove	CNN, NB, SVM e MLP	**97%**
Rosa 2019	Texto	**Remoção de stopword, caracteres estranhos e "Noise Corpus". Stemming.**	Word Embedding não especificado	CNN e RNA	**96%**
Almeida 2021	Texto	**Remoção de dados com conteúdo vazio, acentos e caracteres especiais, stopwords e números e pontuações**	Bag of words TF - IDF	SVM e NB	**80%**
Souza 2020	Texto + Metadados	**Classificação gramatical, polaridade e emoção. Extração de símbolos irrelevantes**	Não especificado	NB, GB, AB, SVM e KNN	**92,53%**

2.2. Base de dados

Para o desenvolvimento do presente trabalho foi adotado a base de dados *Fake.Br Corpus*. Na qual foi construída por pesquisadores da Universidade de São Paulo (USP) [Monteiro et al. 2018]. A base de dados disponibiliza precisamente 7.200 artigos de notícias escritos em português, juntamente com seus respectivos metadados contendo 25 informações sobre os artigos, tais como autor, data de publicação e número de visualizações.

Vale ressaltar que, dos 7.200 artigos, exatamente 3.600 notícias são consideradas verídicas e 3.600 são consideradas falsas. Destaca-se que, estes dados foram coletados dentro de um período de dois anos, de janeiro de 2016 a janeiro de 2018. O conteúdo das notícias da base *Fake.br* são divididos em 6 categorias, sendo estas: política, TV e celebridades, sociedade e notícias diárias, ciência e tecnologia, economia e religião.

2.3. Construção dos cenários

Com base nos passos realizados pelos autores no processo de construção dos modelos de detecção de *fake news* e nos estudos dispostos na literatura, a presente pesquisa realiza a combinação entre estes processos focando em uma etapa por vez, buscando analisar o comportamento dos resultados. Ou seja, é tomado como foco a etapa de transformação e são analisadas as tomadas de decisão adotadas pelos autores acerca da construção dos algoritmos. A seguir é explicado as etapas estudadas e testadas no presente estudo.

2.3.1. Pré-processamento

O pré-processamento dos dados tem como objetivo padronizar os textos, no caso, remover *stopwords*, *tokenizar* os textos e eliminar caracteres irreconhecíveis pelos algoritmos. A etapa de pré-processamento dos textos segue os passos descrito a seguir:

- **Remoção de ruídos e caracteres estranhos:** Remoção de caracteres especiais ou em formato irreconhecível pelos algoritmos, assim como aqueles que não contribuem para o seu treinamento. Neste sentido, numerais, *urls*, caracteres de pontuação e outros caracteres especiais foram eliminados.
- **Remoção de *stopwords*:** Palavras consideradas comuns e, portanto, irrelevantes para a representação de um texto. São normalmente pertencentes às classes gramaticais de pronomes, preposições, artigos, advérbios e conjunções.
- ***Tokenização*:** também conhecida como segmentação de palavras, a *tokenização* consiste na quebra na sequência de caracteres do texto localizando o limite de cada palavra. As partes resultantes deste processo são denominadas de *tokens*.

2.3.2. Transformação

Com o objetivo de facilitar o processamento dos textos pelos algoritmos de AM, é necessário que os textos dos artigos sejam convertidos do formato textual para uma representação numérica. Para isso os textos são submetidos a um processo de vetorização, que consiste em converter cada termo contido no texto em um valor numérico. Desta forma, cada texto é representado por um vetor contendo todos os valores numéricos que representam cada um dos termos contidos nele. Com os textos vetorizados, o processo segue em utilizar a técnica denominada *word embedding*, que converte os textos em vetores multidimensionais, de forma a manter os seus valores tanto sintáticos quanto morfológicos.

2.3.3. Treinamento

Após a etapa de transformação, inicia-se a etapa de treinamento dos algoritmos. Etapa esta em que os algoritmos ”aprendem”a partir dos dados de treinamento e ajustam seus parâmetros para realizar tarefas específicas, como classificação ou regressão. Desta forma, para o processamento dos textos, foi implementada uma CNN e SVM para os textos e para os metadados foram aplicados SVM, NB e MLP. Haja vista que, ambos são comumente aplicados nos trabalhos relatados.

O CNN é um tipo especializado de rede neural amplamente utilizada em tarefas de visão computacional. No entanto, também têm sido aplicadas no processamento de texto. Sendo esta composta por camadas de convolução, que são capazes de extrair características relevantes de forma automática a partir dos dados. O SVM por sua vez, são algoritmos de aprendizado supervisionado que busca encontrar o hiperplano ótimo que melhor separa as classes no espaço de características. Essa abordagem é particularmente útil quando os dados são linearmente separáveis, mas também pode ser estendida para casos não lineares. O NB usa o teorema de Bayes para calcular a probabilidade de que uma determinada amostra pertença a uma classe específica, dadas as características dessa amostra. Já a MLP, é composta por camadas de neurônios conectados, incluindo camadas ocultas que aprendem representações complexas dos dados, sendo capaz de aprender relações não lineares.

3. Resultados

Na presente seção são discutidos os resultados alcançados nos testes realizados. Tendo em vista avaliar o desempenho dos modelos treinados foi selecionado a métricas de avaliação da acurácia. Este critério foi adotado tendo em vista o tamanho da base de dados que está sendo trabalhada.

Como mencionado, o treinamento utilizado foi o CNN e SVM para textos, haja vista que o objetivo consiste em analisar as decisões na criação do algoritmo. É importante relatar que antes de inicializar o desenvolvimento da proposta foi realizado um estudo na base de dados. Como supramencionado este artigo faz uso tanto de textos como de metadados.

No que diz respeito especificamente aos metadados, foi necessário selecionar as informações com maior relevância para o estudo em questão. Para isso, utilizou-se a técnica de *Backward Feature Elimination*, que consiste em iterativamente remover uma coluna por vez do conjunto completo de metadados, reavaliando o modelo a cada etapa. As colunas removidas são aquelas que não causam uma piora significativa no desempenho do modelo ou não contribuem de forma relevante para a análise, resultando em um subconjunto mais conciso e informativo de metadados para a pesquisa.

Os primeiros testes tiveram como foco o processo de transformação, sendo assim, tanto a base de dados, como o nível de análise e o pré-processamento se mantiveram os mesmos para todas as combinações. Neste ponto é importante ressaltar que não foi possível replicar as combinações dos autores [Rosa et al. 2019] e [de Souza et al. 2020] uma vez que, o primeiro utilizou um *embedding* o próprio o qual não está disponível, enquanto o segundo não especificou qual foi aplicado.

Portanto, diante das combinações realizadas, como pode ser visualizada na Tabela2, foi notório que a diferença entre os resultados é mínima, já que todos alcançaram uma média de 97% de acurácia, porém dentre as combinações o que mais se sobressaiu foi o transformador *Word2Vec* com 97,7%. De posse dos resultados constata-se que o nível de influência do transformador no algoritmo não é tão forte quanto outros fatores.

Tabela 2. Resultados do primeiro experimento.

	Transformação	Treinamento	Acurácia
Teste 1	**Glove**	CNN	**97,0%**
Teste 3	**Wang2Vec**	SVM	**96,4%**
Teste 4	**Word2Vec**	SVM	**96,8%**
Teste 5	**TF-IDF**	SVM	**96,8%**
Teste 6	**FASTEXT**	SVM	**96,6%**

Os demais testes tiveram como foco o pré-processamento, porém é consensual que os processos de remoção de caracteres estranhos e tokenização são etapas fundamentais para a preparação dos textos para a análise. Sendo assim, a fim de analisar o impacto do pré-processamento foram realizados testes comparativos a respeito da remoção ou manutenção de *stopwords*.

Como pode ser visualizado na Tabela 3 sem a remoção de *stopwords* foi o cenário que apresentou o melhor resultado com 97% de acurácia, aplicando o TF-IDF.Frisa-se

que os demais resultados foram semelhantes, sendo que a combinação do Glove sem a remoção de *stopwords* apresentou menor acurácia. Tendo em vista os resultados obtidos constatou-se que, especificamente para o cenário sem a remoção de *stopwords* para o TF-IDF a remoção de *stopwords* pode levar a perda de informações contextuais já que elas fornecem estrutura gramatical e conectividade entre as demais palavras. Isso pode resultar em uma representação menos rica dos documentos, levando a uma perda de informações relevantes.

Tabela 3. Resultados do segundo experimento.

	Pré - Processamento	Transformação	Treinamento	Acurácia
Teste 1	Com Stopwords	**Glove**	CNN	**96,0%**
Teste 2	Sem Stopword	**Glove**	CNN	**96,3%**
Teste 3	Com Stopword	**TF-IDF**	SVM	**97,0%**
Teste 4	Sem Stopword	**TF-IDF**	SVM	**96,8%**
Teste 5	Com Stopword	**BOW**	SVM	**96,9%**
Teste 6	Sem Stopword	**BOW**	SVM	**96,8%**

No que diz respeito aos metadados os resultados foram bem próximos, porém o SVM apresentou uma melhor adequação aos dados com 97,6% de acurácia, seguido do MLP com 97,4% e por fim o NB com 95,7%.

4. Considerações finais

Este trabalho propõe uma avaliação de desempenho das etapas do processo de extração de conhecimento para detecção de *fake news* em português. Realizando uma análise do processo de construção dos algoritmos, investigando o impacto das tomadas de decisão no processo metodológico no resultado final, proporcionando assim, encontrar a melhor combinação na literatura atual para a detecção de notícias falsas escritas em português. Para chegar a este objetivo uma série de estudos foi realizada, como discorrido ao longo deste artigo. No qual, foi possível identificar características de notícias falsas e verdadeiras, assim como a carência de estudos na área de detecção de *fake news* em português.

Analisando as combinações foi perceptível que os maiores impactos nos resultados foram verificados quando o foco se encontrou no pré-processamento, mais especificamente com relação às *stopwords*. No caso sem a remoção de *stopwords* foi o cenário que apresentou o melhor resultado com 97% de acurácia, aplicando o *TF-IDF*. Em contrapartida, a combinação do *Glove* sem a remoção de *stopwords* apresentou menor acurácia.

Como trabalhos futuros, pretende-se expandir as combinações, no caso, aplicar técnicas utilizadas na língua inglesa e ver o comportamento no português. Destaca-se que pretende-se refazer os testes em outras bases de dados em português já consolidadas.

Referências

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31:211–236.

Almeida, L., Fuzaro, V., Nieto, F. V., and Santana, A. (2021). Identificação de "fake news" no contexto político brasileiro: uma abordagem computacional. In *Anais do II Workshop sobre as Implicações da Computação na Sociedade*, pages 78–89, Porto Alegre, RS, Brasil. SBC.

Aro, J. (2016). The cyberspace war: Propaganda and trolling as warfare tools. *European View*, 15(1):121–132.

Carvalho, C., Klagge, N., and Moench, E. (2011). The persistent effects of a false news shock. *Journal of Empirical Finance*, 18(4):597–615.

DataReportal (2022). DIGITAL 2023: GLOBAL OVERVIEW REPORT. Disponível em: `https://datareportal.com/reports/digital-2023-global-overview-report`. Acesso em: 02 Maio 2023.

de Souza, M. P., da Silva, F. R. M., Freire, P. M. S., and Goldschmidt, R. R. (2020). A linguistic-based method that combines polarity, emotion and grammatical characteristics to detect fake news in portuguese. *Brazilian Symposium on Multimedia and the Web*.

Fuller, C. M., Biros, D. P., and Wilson, R. L. (2009). Decision support for determining veracity via linguistic-based cues. *Decision Support Systems*, 46(3):695–703. Wireless in the Healthcare.

Gonzalez, M. and Lima, V. L. S. d. (2003). Recuperação de informação e processamento da linguagem natural. *XXIII Congresso da Sociedade Brasileira de Computação*.

Kogan, S., Moskowitz, T. J., and Niessner, M. (2019). Fake news: Evidence from financial markets. *Available at SSRN*, 3237763.

Kumar, S., Kumar, S., Yadav, P., and Bagri, M. (2021). A survey on analysis of fake news detection techniques.

Monteiro, R. A., Santos, R. L. S., Pardo, T. A. S., de Almeida, T. A., Ruiz, E. E. S., and Vale, O. A. (2018). Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In *Computational Processing of the Portuguese Language*, pages 324–334. Springer International Publishing.

Reuters, I. and of Oxford, U. (2021). Reuters Institute Digital News Report 2021. Disponível em: `https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2021-06/Digital_News_Report_2021_FINAL.pdf`. Acesso em: 18 Maio 2022.

Rosa, M. d. A., Amorim, H. C., Gomes, R. M., and Santos, B. A. (2019). Detecçao de fake news: uma abordagem utilizando redes neurais convolucionais.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., and Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Trans. Intell. Syst. Technol.*, 10(3).

Sousa, F., Barbosa, A., Oliveira, C., and Braga, R. (2022). Detecção de fake news em língua portuguesa combinando redes neurais convolucionais e algoritmos de aprendizagem de máquina. *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*.