DETECTORES DE FAKE NEWS: ANÁLISE E IMPLEMENTAÇÃO

André Luiz Neilsen Carneiro De Castro (FHO – Fundação Hermínio Ometto)
andreluizcarneiro@outlook.com.br
Tiago Pereira Remédio (FHO – Fundação Hermínio Ometto) remedio@fho.edu.br

Resumo

Com a expansão dos meios de comunicação, as notícias transcorrem de forma dinâmica, e estão cada vez mais disponíveis na internet, seja em portais jornalísticos ou nas diversas redes sociais. A maior parte das notícias são pautadas em informações verificadas, seguindo o devido rigor jornalístico de verificação das fontes. No entanto, a facilidade da internet possibilitou que *fake news* (notícias falsas) sejam criadas, com a intenção de enganar leitores que estão atrás do consumo de informações. A partir de uma aplicação *web*, o objetivo deste trabalho é apontar uma porcentagem da notícia ser verdadeira (por meio da implementação de um *web crawler*) utilizando um algoritmo de abordagem própria para detecção no processamento de texto e na identificação de características de escrita. Por meio dos resultados o sistema apresentou-se como uma boa ferramenta de apoio para os usuários, contribuindo com que o ecossistema de notícias mantenha sua autenticidade. Dentre algumas tecnologias utilizadas estão: HTML, CSS, PHP, PYTHON e o gerenciador de base dados MySQL. O sistema foi desenvolvido pensando em um ambiente visual responsivo e intuitivo.

Palavras-Chaves: detector de *fake news*, desenvolvimento *web*, *web crawler*

1. Introdução

Com a popularização da internet e facilidade de acesso às informações possibilitadas pelas mídias sociais, redes sociais e aplicativos de mensagens, a rápida disseminação de notícias falsas vem preocupando toda uma sociedade em relação a confiabilidade de informações (COSTA, 2019). A ampla divulgação de notícias falsas pode ter um sério impacto negativo nos indivíduos, influenciando negativamente em vários setores da sociedade, como política, saúde e segurança, podendo quebrar o equilíbrio de autenticidade do ecossistema de notícias, persuadindo intencionalmente os consumidores para que aceitem crenças tendenciosas ou serem utilizadas para transmitir informações políticas falsas (SHU et al., 2017).

Há um interesse crescente e contínuo no desenvolvimento de ferramentas automatizadas para identificar informações incorretas e tendenciosas (GRUPPI et al., 2018). No entanto, detectar

notícias falsas nas redes sociais apresentam várias novidades e problemas de pesquisa desafiadores para cada tipo de notícia analisada (SHU et al., 2017).

No que se refere ao tratamento e detecção de notícias falsas, a aplicação de diversas tecnologias acaba se tornando comum em prol de buscar uma assertividade cada vez maior e diminuir os erros proeminentes. A base delas, e imprescindível, é o *web-crawler*, um mecanismo de busca que detecta e varre dados automaticamente nas páginas da internet. A utilização de *datasets* tem se tornado cada vez mais comum, que por sua vez, possui notícias já rotuladas com suas deveras características de credibilidade, tendo passadas por jornalistas, profissionais e especialistas. A utilização de técnicas de *word embedding* (transformação de palavras em vetores contendo informações numéricas) em conjunto com *machine learning* (método de análise de dados que automatiza a construção de modelos analíticos) e algoritmos de aprendizado de máquina são as tecnologias que obtiveram resultados mais satisfatórios.

O objetivo deste trabalho é unificar alguns desses métodos e trazer um algoritmo classificatório de linguagem natural, atribuindo um fator de credibilidade para as notícias inseridas pelo usuário após terem sido levantadas pelo *crawler*.

2. Revisão bibliográfica

2.1. Comunicação atual

O trabalho de divulgação de notícias sobre política pela mídia é profundamente discutido e questionado há décadas. Os domínios dos meios de comunicação representam uma ferramenta poderosa capaz de influenciar e manipular a opinião da sociedade, alterando a percepção dos acontecimentos. (SILVA et al., 2019).

O ecossistema da mídia de notícias mudou ao longo do tempo de jornal para rádio/televisão e, recentemente, notícias online e redes sociais. (SHU et al., 2017). A ampliação do consumo de notícias por sites de redes sociais impulsiona um novo tipo de concorrência com os meios de comunicações tradicionais. (DELMAZO et al., 2018).

Associado a importância de textos de notícias e seu compartilhamento das mesmas em redes sociais, vem o compartilhamento exponencial de notícias falsas, criadas com intenção desonesta de enganar os consumidores. A quantidade de eventos e debates acerca deste fenômeno que vem sendo chamado de *fake news* tornou-se recorrente, mobilizando pesquisadores de diversas

áreas com o objetivo de entender, avaliar o seu impacto e combatê-las. (MONTEIRO et al., 2019).

2.2. *Fake news*

Diante da facilidade de criação e compartilhamento de informações, as mídias sociais tornaram-se um meio prático para disseminação de *fake news*, sendo a principal fonte de informação desse século, causando grandes transtornos para a sociedade e trazendo grandes impactos negativos, sendo identificada como uma ameaça global (ALLCOTT; GENTZKOW, 2017). As motivações subjacentes dos criadores e a disseminação de informações falsas variam, e a desinformação pode existir em várias formas (BOBERG et al., 2019).

Mesmo que alguns tipos de notícias apresentem características semelhantes, elas podem conter propriedades e atributos diferentes (COSTA, 2019). No entanto, podemos identificar diversos tipos distintos de *fake news* que repercutem intencionalmente com a finalidade de enganar os leitores. Vão desde a sátira ou paródia sobre manchetes enganosas, conteúdo enganoso e informações contextuais falsas até conteúdo impostor (com fonte tendenciosa), conteúdo manipulado, conteúdo totalmente fabricado e falsa conexão. Tais tipos estão presentes na Figura 1. (WARDLE et al., 2017). As *fake news* buscam ser emocionalmente apelativas, dificultando que o leitor desse conteúdo tenha uma reação racional sobre aquilo que acaba de ler, muitas vezes almejando obter ganhos financeiros ou políticos. A reação emocional induz um comportamento impulsivo do leitor, que também ajuda a explicar o alto índice de compartilhamento dessas notícias (SIVEK, 2018). As *fake news* estão presentes na crise de confiança dos leitores nos veículos tradicionais de comunicação. (DELMAZO et al., 2018).

Figura 1 – Tipos Distintos De Fake News

Fonte: Adaptado de WARDLE (2017)

Existem alguns estudos que apontam que além das emoções e das outras características citadas, existem outros atributos únicos que ajudam propagar notícias falsas nas redes sociais, que também se aplicam as mídias tradicionais. (SHU et al., 2017).

a) Contas maliciosas nas redes sociais para propagandas, que surge através do baixo custo de criação de mídia social, incentivando contas de usuários maliciosos, como usuários ciborgues, comportamentos inautênticos, contas falsas, *trolls* e *bots* sociais que são projetados maliciosamente com o propósito de criar danos.
b) *Echo chamber effect*, que se tratam de usuários seguindo pessoas com ideias semelhantes e que promovem e favorecem suas próprias narrativas, polarizando suas opiniões resultando na reverberação da notícia.
c) Fatores psicológicos, onde as pessoas são mais propensas a aceitar uma fonte como confiável, principalmente quando há poucas informações sobre aquela notícia.

As *fake news* implicam em teorias da conspiração e fatos alternativos, fornecendo conteúdos, argumentações e referências para seus seguidores, distanciando grupos de pessoas cada vez mais dos fatos e centralizando a sociedade a ponto de que grupos distintos acreditem em falsas notícias criadas (MANJOO, 2016). Todos nós desempenhamos um papel crucial neste ecossistema. Sempre que aceitamos informações sem verificar duas vezes ou compartilhamos

uma postagem, imagem ou vídeo antes de verificá-los, estamos aumentando a proliferação de notícias falsas. (WARDLE, 2017).

Uma das formas de combater *fake news*, seria combater a desinformação com mais informação verdadeira, se aproximando do leitor através das redes sociais e capacitando-o de identificar notícias falsas, explicando a importância de verificá-las antes de qualquer compartilhamento. Além disso, detectores de *fake news* crescem exponencialmente com a finalidade de ajudar o leitor, porém características e desafios únicos tornam os algoritmos de detecção existentes ineficazes ou não aplicáveis.

> "Detectar Notícias falsas se tornaram cada vez mais importantes e podem beneficiar indivíduos e até mesmo nossa sociedade em muitos aspectos. Primeiro, as pessoas estarão bem informadas sobre eventos e notícias e suas atividades políticas e sociais não serão mal orientadas. Em segundo lugar, apesar de algumas iniciativas recentes de alguns provedores de mídia social, como Facebook, não há detecção sistemática de notícias falsas por plataformas de mídia social. Terceiro, identificar as notícias falsas são um passo em direção a incentivos financeiros que estimulem os divulgadores que administram seus "negócios". Por exemplo, o Google tenta parar de fornecer seus serviços de anúncios a sites de notícias falsas" (KARIMI et al., 2018).

2.3. Inteligência Artificial

Acompanhado dos métodos de detecção de notícias, estão ferramentas como *web crawler*, que coleta informações da *web* em grande escala, possuindo um módulo de filtro que desempenha papéis importantes em um rastreador da *web*, um componente central de um mecanismo de pesquisa. Os desempenhos de estabilidade e paralelismo do algoritmo são verificados pelos experimentos para *sites* que lidam com um grande número de páginas da *web*. Resultados mostram que algoritmos mesclados com o *crawler* podem alcançar desempenhos satisfatórios na detecção de *fake news* (HUI-CHANG et al., 2009).

Com a intenção de melhorarem a acurácia do detector de notícias e devido à confiança nas fontes convencionais de notícias estarem cada vez menores, os pesquisadores estão trabalhando em maneiras de dar às pessoas mais *insights* sobre acreditarem no que leem. Os pesquisadores têm testado ferramentas de inteligência artificial (IA) que combinados com o *web crawler*

podem ajudar a filtrar notícias legítimas com uma maior eficácia (HOROWITZ, 2021). Com esse objetivo, floresceram mecanismos como *machine learning*, sendo muito usado para processar grandes quantidades de dados de treinamento a fim de classificar ou identificar dados novos, semelhantes aos que foram treinados (BASHEER, et al, 2000). A abordagem de *machine learning* com algoritmos através dos anos incluindo árvore de aprendizado, programação lógica indutiva, aprendizado profundo, agrupamento, aprendizado reforçado, redes Bayesianas, entre outros, trouxeram resultados significativos, cada um apresentando um grau de imprecisão e se destacando de formas diferentes devido aos desafios encontrados na detecção de *fake news*.

Quando se trata de *machine learning*, antes de detectar notícias falsas é necessário "treinar" a máquina usando uma quantidade grande de dados e algoritmos que dão a ela a habilidade de aprender como executar a tarefa, uma estruturação de *bag of words* (representações de texto que descreve a ocorrência de palavras em um documento) e *word embedding*, que consiste em uma representação de palavras que permite que palavras com significados semelhantes tenham uma representação semelhante, devem ser bem elaboradas. Elas são uma representação distribuída de texto que contribuem para o desempenho de métodos como o de aprendizado profundo em problemas desafiadores de processamento de linguagem natural (BROWNLEE, 2017).

As ferramentas de IA são ótimas para lidar com grandes quantidades de informações em alta velocidade, mas carecem da análise diferenciada que um jornalista ou que um verificador de fatos pode fornecer. Apesar da eficácia dos algoritmos, a intervenção humana é necessária (HOROWITZ, 2021).

2.4. Tecnologia na sociedade e integração com notícias

Atualmente conectados 24 horas por dia, podemos acompanhar a qualquer instante tudo que ocorre do outro lado do mundo. Dessa forma, a inovação tecnológica nos proporciona evolução e revolução. Quem não se adaptar com essa contínua mudança tecnológica fica desatualizado e fora do contexto social (FERREIRA, 2017).

Após décadas de relativa estabilidade nos mercados de mídia, o século XXI produziu condições que mudaram rapidamente, caracterizadas por mudanças tecnológicas e altos níveis de insegurança (GADE et al, 2015). Nesse ambiente, as empresas de mídia de notícias tornam-se mais flexíveis e buscam uma mudança contínua. Apesar das inovações tecnológicas criarem um potencial para novos produtos, perseguir essas oportunidades é incômodo para os padrões

estabelecidos de trabalho e processos de organização, promovendo um risco considerável. Indústrias enfrentam a exigência de mudança tecnológica para entender o impacto das novas tecnologias na organização a fim de explorar as oportunidades que as novas tecnologias podem oferecer. A inovação tecnológica muitas vezes leva as organizações a repensarem a forma como elas são estruturadas. (GALBRAITH, 1994).

2.5. Trabalhos relacionados

Com a expansão dos meios de comunicação e com o avanço da tecnologia ao passar dos anos, as notícias transcorrem de forma dinâmica, trazendo desafios em detectar notícias falsas devido ao grande grau de complexidade.

Kai Shu, at al. (2017), apresentam reflexões a respeito de disseminações de informações onde levam pessoas a buscar e consumir notícias falsas nas redes sociais. Retratando o impacto negativo em consumi-las e os desafios encontrados ao utilizar algoritmos nas detecções. A abordagem teórica aprofundada na caracterização de notícias falsas e as métricas e perspectivas dos algoritmos existentes retratadas no artigo, fornecem uma visão ampla sobre as *fake news*.

Wang Hui-Chang, Ruan Shu-Hua e Tang Qi-Jie (2009), buscam encontrar desempenhos satisfatórios por meio de um *web crawler* que desempenha papéis importantes rastreando as notícias e através de um algoritmo de filtro realiza o tratamento. Os resultados obtidos pelo experimento mostraram que o algoritmo pode alcançar desempenhos efetivos por meio de ajustes razoáveis combinado com o *web crawler*.

Verónica Pérez-Rosas, at al. (2017), abordam meios de automação e identificação de notícias falsas utilizando métodos de classificação que dependem de recursos que representam propriedades de legibilidade do texto, como combinação léxica, sintática e informações semânticas. Apesar de complexo, o processamento de linguagem natural possui uma área muito vasta e envolve diversos conhecimentos, onde técnicas capazes de alcançar precisões comparáveis à capacidade humana de detecções textuais podem ser adotadas de diversas formas.

3. Metodologia

Este trabalho utiliza o método de pesquisa experimental, onde serão abordados o tema de *fake news* e sua detecção utilizando análises e sínteses experimentais. O estudo terá fonte de pesquisa

primária (com artigos, relatórios e projetos) e fontes secundárias. A pesquisa tem caráter qualitativo (apontando a análise e conceito sobre o estudo). O trabalho também conta com uma pesquisa bibliográfica e estudo de campo (capturando e analisando dados disponíveis publicamente na internet).

O desenvolvimento do trabalho envolve a linguagem de programação *python*, junto com a linguagem de programação *web PHP* que será utilizada para fornecer as informações para o usuário. As pesquisas realizadas pelo usuário serão armazenadas no banco de dados MySQL, onde será feito um levantamento de precisão do sistema.

Primeiramente, cria-se a parte de *backend* do sistema, onde irá gerir todos os procedimentos textuais e analíticos responsáveis para designar a informação final para o usuário. Em seguida, realiza-se a criação do *frontend* onde será utilizado HTML e CSS, responsável pela experiência e interação do usuário na aplicação *web.* Após a aplicação dos procedimentos, realizam-se os testes e análises das informações obtidas, para garantir a experiência do usuário, a eficiência do sistema e obter conclusões a respeito dos resultados da aplicação.

4. Desenvolvimento

Visando desenvolver o projeto utilizando uma abordagem própria, foram levantadas algumas informações relevantes em relação ao ecossistema de notícias no Brasil e ao redor do mundo e como tecnicamente os algoritmos estão tratando os problemas atuais dessa proliferação de notícias falsas.

Também foi pensado quais seriam as melhores tecnologias a serem utilizadas para análises textuais e qual seria o melhor caminho para fornecer para o usuário um resultado relevante, após realizar o tratamento de qualquer tipo de notícia que venha ser pesquisada nos mecanismos de busca.

4.1. Páginas *web*

A *web* possui diversas ramificações, uma delas são as páginas *web*, também conhecidas como páginas eletrônicas, onde são armazenadas em um computador conhecido como servidor da *web*. Para que a página da *web* seja exibida nesse ou em outro computador, ela deve ser acessada e interpretada por um programa específico, denominado *software cliente*. A rede de internet é

feita dessa forma, com computadores interligados formando essa rede e hospedando páginas e sites que são acessados por usuários ao redor do mundo.

Ao iniciar o desenvolvimento, mecanismos de buscas (sistemas online encarregados de pesquisar arquivos armazenados em servidores da internet) foram de extrema importância pois trazem páginas da *web* e cada um utiliza um critério específico ao realizar a busca do que foi inserido pelo usuário, podendo utilizar em sua composição *spiders* ou *crawlers*, diretórios ou catálogos de busca, mecanismos de *metabusca*, mecanismos híbridos, entre outros. Nenhum dos mecanismos de buscas divulgam detalhadamente sobre os critérios adotados nas suas classificações, já que a tecnologia utilizada em cada um é o diferencial em relação a concorrência.

Na sua grande maioria, os buscadores trabalham com *search engine optimization*, onde tal técnica, permite que classifiquem uma fonte de conteúdo relevante sobre determinado assunto. Algumas características são necessárias para essa classificação, são elas: estruturação, sem *links* quebrados, um código sucinto, imagens formatadas, título, a ocorrência das palavras no conteúdo da página e procuram ver a relação das palavras da pesquisa com a página que está sendo buscada. Além disso, utilizam um conjunto de táticas que visam adquirir *links* qualificados e relevantes de outros sites (*link building*).

Uma vez que os próprios mecanismos de busca realizam uma classificação de notícias refinada e levando em consideração que a maioria dos internautas consomem as notícias apenas da primeira página de busca, um *web crawler* foi criado para capturar essas notícias que são geradas na primeira página de pesquisa.

4.2. *Web crawler*

Com intuito de varrer as informações obtidas nas principais páginas *web* e armazená-las para realizar os tratamentos, foi desenvolvido um *web crawler*. O *crawling* é realizado no *Google* (maior mecanismo de busca de todos os tempos), *DuckDuckGo* (mecanismo que utiliza informações de origem *crowdsourcing* para melhorar a relevância dos resultados) e *Yahoo* (terceiro mecanismo de busca mais usado). As informações obtidas e armazenadas que foram extraídas pelo *crawler*, foram o título e o *meta-description* (resumo da descrição da página pesquisada). Com essas informações levantadas que foram armazenadas em vetores, será trabalhado todo o projeto.

4.3. Analisador e detector de notícias

Após o resumo e o título terem sido capturados pelo *crawler*, um algoritmo mais "humano" foi pensado, para que tentasse retratar o procedimento que os seres humanos normalmente realizam nos mecanismos de busca para identificar se uma notícia é falsa ou não, através da análise do processamento textual. O processamento textual é um dos aspectos essenciais para a análise e detecção de notícias falsas, dessa forma será utilizado a linguagem *python* para o desenvolvimento, onde possui a biblioteca NLTK, uma tecnologia capaz de ajudar o computador a entender a linguagem natural humana, possuindo recursos amplos para o processamento de texto.

Procurando reconhecer a notícia pesquisada pelo usuário nos nossos dados extraídos pelo *crawler*, um algoritmo que calcula a similaridade das palavras com base na distância e no comprimento foi utilizado. O algoritmo de *Levenshtein* utiliza uma métrica de *string* para medir a diferença entre duas palavras utilizando o número mínimo de operações necessárias para transformar uma *string* na outra, sendo muito utilizado para determinar quão semelhante as *strings* são e apresenta uma porcentagem de semelhança quando comparadas. Dessa forma, foi utilizado para reconhecer os títulos e os resumos obtidos pelo *crawler* com a informação inserida pelo usuário, nos trazendo um fator positivo de que a notícia obtida possa se tratar daquela que procuramos, ajudando a eliminar falsos positivos.

Um obstáculo ao utilizar esse algoritmo para encontrarmos notícias referentes ao que foi pesquisado pelo usuário, é quando em uma determinada sentença encontrada pelo *crawler* tem variação do verbo ao longo das buscas por um sinônimo, dessa forma a porcentagem de similaridade fornecida pelo algoritmo diminui, aparentando ser uma notícia que não pertence ao assunto pesquisado (Figura 2). Dessa forma para uma maior apuração de informações textuais levando em consideração que a língua portuguesa possui mais de 5000 verbos, foi criado um dicionário de sinônimos, que percorre todas as notícias encontradas pelo *crawler* por uma segunda vez substituindo os verbos por um sinônimo daquele verbo, buscando aumentar o grau de similaridade.

Figura 2 – Aplicação do Algoritmo de Levenshtein

Sentença utilizada na pesquisa: Agnaldo Timóteo morreu **Sentença encontrada no mecanismo de busca:** Agnaldo Timóteo morreu **Distância de Levenshtein:** 0 **Semelhança:** 100%	**Sentença utilizada na pesquisa:** Agnaldo Timóteo morreu **Sentença encontrada no mecanismo de busca:** Agnaldo Timóteo faleceu **Distância de Levenshtein:** 5 **Semelhança:** 80%

Fonte: Criado pelo autor

No desenvolvimento, partindo da premissa de que iria ser trabalhado a sentença que foi inserida na pesquisa pelo usuário e que não estaríamos utilizando uma *bag of words*, foi utilizado uma técnica similar ao preparamento de texto para *machine learning*, onde consiste em analisar a ocorrência daquela palavra. Portanto, quando o usuário insere uma determinada notícia verificamos a ocorrência dela no título e no resumo da notícia. Se todas as palavras existirem no que foi levantado pelo *crawler*, quer dizer que existe a chance da notícia ser verdadeira, ou seja, o fato existe. Quanto menor a incidência das palavras, menor a chance da notícia ser verdadeira, implicando que a pesquisa pode se tratar de uma *fake news*.

Inicialmente para calcular a porcentagem que indica que a notícia é verdadeira, foram considerados dois critérios: 1) analisar quantas palavras da pesquisa aparecem na notícia, compondo 50% do total do cálculo; 2) verificar se o termo inserido pelo usuário existe na notícia, compondo o restante dos 50%. Contudo, ao realizar os testes, eram apresentadas muitas imprecisões na porcentagem, onde em muitos casos um critério poderia alavancar a porcentagem e outro acarretaria em diminuir a porcentagem, trazendo muitas imprecisões.

Pensando em mesclar os dois métodos utilizados anteriormente para diminuir imprecisões e trazer um equilíbrio na porcentagem fornecida para o usuário, o algoritmo foi trabalhado comparando as palavras do termo pesquisado em cada notícia e contabilizando o número de ocorrências e a distância entre elas. Se todas as palavras do termo estão presentes na notícia, classificamos que ocorre 100% de ocorrência. Em seguida o algoritmo avalia quantas palavras aparecem entre o termo pesquisado, sendo assim, caso nenhuma palavra desconhecida apareça, classificamos que a ocorrência do termo é totalmente consistente, ou seja, 100%. Cada palavra que aparece entre o termo pesquisado, é descontado 7,5% do total da porcentagem de consistência. A técnica de calcular a consistência do termo pesquisado foi utilizada pois ela

ajuda identificar se as palavras encontradas estão no mesmo contexto, eliminando falsos positivos. Desse modo, o resultado final que é apresentado para o usuário é a porcentagem da ocorrência das palavras multiplicado pela porcentagem da consistência do termo pesquisado.

Calculando a porcentagem de todas as notícias temos a ocorrência das palavras (*O*), a porcentagem de consistência (*C*) onde começa com 100% e vai diminuindo em 7,5% a cada palavra encontrada da qual não pertence a sentença inserida pelo usuário. Para porcentagem de cada notícia (*PN*), temos $PN = O \times ((100 - C) / 100)$. Calcula-se, então, o total de notícias, onde possuímos o número de notícias (*N)*, soma de porcentagem de todas as notícias (*PNS*), a porcentagem média final de todas as notícias (*PF*). A equação final fica, então, $PF = PNS/N$.

Utilizando como exemplo a sentença pesquisada na Figura 2, onde apresenta o texto "Agnaldo Timóteo Morreu", as notícias encontradas continham 100% das palavras da pesquisa e em todas elas não há o aparecimento de outras palavras entre a sentença pesquisada, apenas a inversão (morre Agnaldo Timóteo). Dessa forma ficará N = 10, O = 100% para todas as notícias, C = 100 para todas as notícias. Logo PN = 100% para todas as notícias, então PNS = 1000%. Finalmente: PF = 1000/10 = 100%.

Buscando tornar o conteúdo processado mais eficiente foi adicionado uma biblioteca chamada *stop words* (conhecido no processamento de linguagem natural, como palavras de parada) onde possui uma lista com mais de 16 idiomas diferentes e que remove todos os artigos encontrados no texto.

Por fim, uma aplicação *web* foi montada, apresentando uma porcentagem daquela notícia pesquisada para o usuário que utiliza como critério as ocorrências encontradas em todas notícias levantadas pelo *crawler*. Caso apareçam palavras nas buscas como *fake news, "*esta notícia é falsa", "esta notícia é tendenciosa", aparecerá um campo de observação no resultado da pesquisa que indica que a notícia pode se tratar de uma *fake news*, isto não afetaria na porcentagem fornecida para o usuário.

4.4. Expondo os resultados da análise

Um segundo critério de classificação de notícias após serem levantadas pelo *crawler* se torna inusual, pois existem fatores benéficos ao utilizarmos os próprios meios de classificação dos mecanismos de busca, um deles acontece quando inserimos uma notícia que provavelmente não

existe, ela nos traz inúmeros falsos positivos, do qual na maioria das vezes trata-se de notícias aleatórias e isso facilita na hora de aplicarmos os critérios de detecções abordados.

Uma forma de classificação tinha sido pensada ao longo do projeto, no qual consistia em dividir os tipos de notícias em pilares, onde se tratariam de tipos de notícias mais pesquisadas nos mecanismos de busca, como: fatos imediatos, entretenimento, conteúdo próximo aos interesses do usuário, identidade do leitor, etc. Todos eles seriam capturados pelo *crawler* por alguns mecanismos de buscas e feito um comparativo com o *site* do qual a incidência daquela notícia poderia ocorrer com mais facilidade. No entanto, como a gama de pesquisa aumentou, também aumentou o grau de complexidade, onde muitas vezes poderia funcionar para uma notícia ou outra, mas a imprecisão obtida de forma geral, era muito maior do que se aplicado no próprio conteúdo levantado pelos mecanismos de busca.

Diante destes fatores, a forma classificativa que os próprios mecanismos de busca já possuem, anulam uma segunda tentativa de classificação de notícias, que poderia gerar um possível retrabalho desnecessário e com mais imprecisões.

5. Resultados

As combinações das técnicas utilizadas nos trouxeram resultados significativos e interessantes para alguns tipos de notícias e, para outros, os resultados foram inconsistentes. Devido não possuir um *dataset* e o tratamento ser diretamente com a notícia que o usuário digitou comparando-a com o que foi obtido pelo *crawler*, a porcentagem fornecida para o usuário traz um resultado interessante para muitos casos analisados.

Buscando melhorar a precisão dos resultados, o novo algoritmo de ocorrências desenvolvido teve uma melhor performance, pois analisa a aproximação das palavras pesquisadas nas notícias. Dessa forma, as notícias são classificadas com maior porcentagem em relação a primeira técnica, pois o algoritmo calcula a consistência do termo pesquisado, isto é, considera as palavras que aparecem entre o termo pesquisado.

Como apresentado na Figura 3, notícias utilizando o novo algoritmo criado comparadas com o algoritmo anterior, mostraram uma melhora na porcentagem fornecida para o usuário para quaisquer tipos de notícias que são pesquisadas.

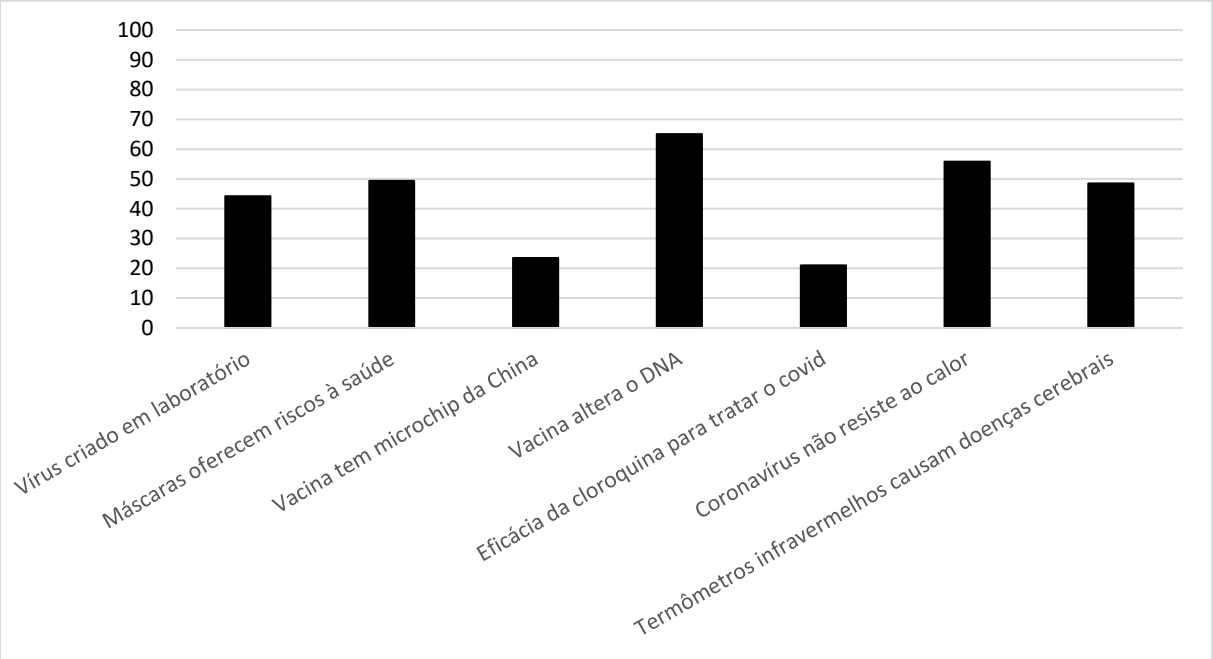
Figura 3 – *Fake news* relacionadas ao Covid-19

Sentença utilizada na pesquisa: Concurso Miss França é processado **Porcentagem obtida no algoritmo anterior:** 90.57% **Porcentagem obtida no novo algoritmo:** 96.97%	**Sentença utilizada na pesquisa:** Marília Mendonça cai de avião **Porcentagem obtida no algoritmo anterior:** 66.67% **Porcentagem obtida no novo algoritmo:** 87.65%	**Sentença utilizada na pesquisa:** Bolsonaro concede medalha a si mesmo **Porcentagem obtida no algoritmo anterior:** 68.23% **Porcentagem obtida no novo algoritmo:** 87.45%
Sentença utilizada na pesquisa: Índia testa míssil intercontinental **Porcentagem obtida no algoritmo anterior:** 56.96% **Porcentagem obtida no novo algoritmo:** 82.05%	**Sentença utilizada na pesquisa:** Sergio Moro planeja disputar a presidência **Porcentagem obtida no algoritmo anterior:** 28.33% **Porcentagem obtida no novo algoritmo:** 46.16%	**Sentença utilizada na pesquisa:** Reajuste no preço da gasolina **Porcentagem obtida no algoritmo anterior:** 68.42% **Porcentagem obtida no novo algoritmo:** 76.50%

Fonte: Criado pelo autor

Para o levantamento de informações específicas em relação as maiores notícias relacionadas a *fake news* dos últimos tempos, testes foram realizados em benefício de obter a precisão do sistema. Começando pelo levantamento de *fake news* relacionados a COVID-19, que, logo após o surgimento do vírus, centenas de notícias falsas e de boatos se espalharam. Um levantamento da porcentagem em relação a algumas notícias relacionadas ao COVID-19 é abordado no gráfico da Figura 4.

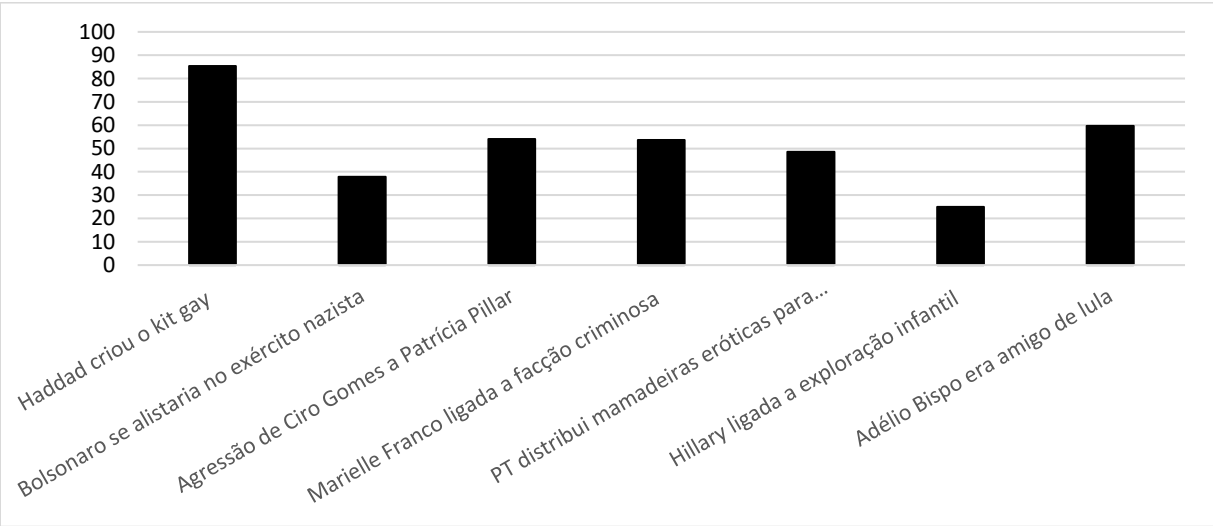
Figura 4 – *Fake news* relacionadas ao Covid-19



Fonte: Criado pelo autor

As informações obtidas e apresentadas no gráfico anterior, apresentaram em média 43.92% chances da notícia pesquisada ser verdadeira e todas acusaram que independente da notícia pesquisada, poderia se tratar de uma *fake news.* No gráfico da Figura 5, obteve-se resultados em relação a algumas *fake news* polêmicas relacionadas a política no Brasil e no mundo entre 2020-2021.

Figura 5 – *Fake news* relacionadas a política

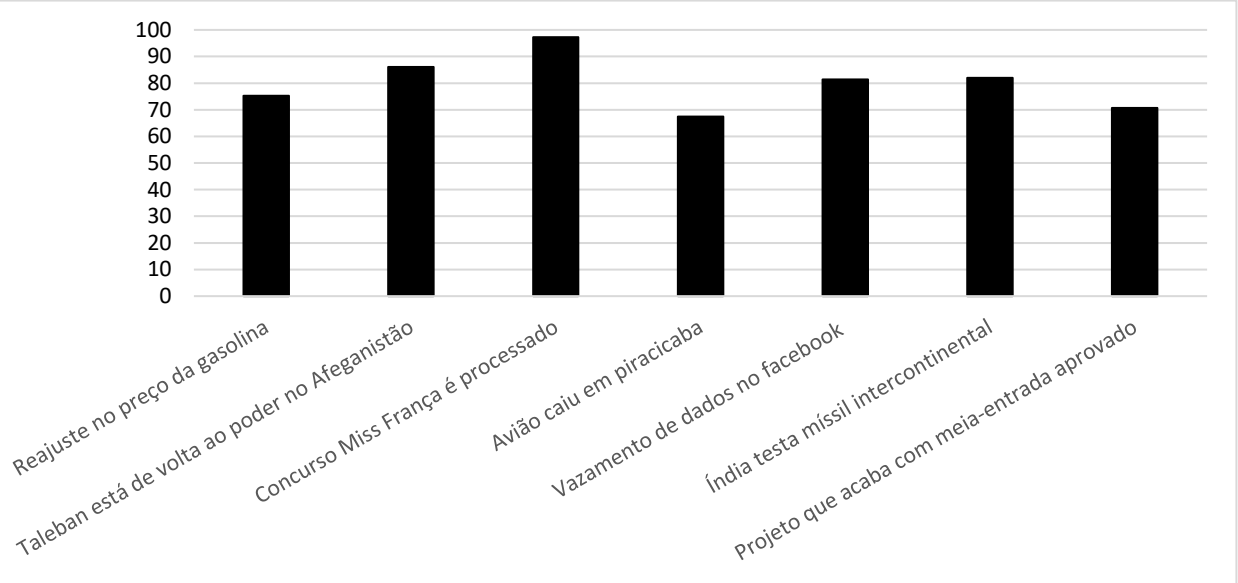


Fonte: Criado pelo autor

As porcentagens obtidas nos resultados apresentados mostram porcentagens com indicativos que todas se tratam de notícias falsas e com o índice abaixo de 60% com exceção da primeira notícia onde envolve o pré-candidato à presidência na época, Haddad, possuindo 85.34% onde indica que possivelmente a notícia é verdadeira, no entanto, esta notícia já foi comprovada se tratar de *fake news* por diversos meios de comunicação. Isso mostra que o algoritmo nem sempre mostra um baixo índice de porcentagem para uma notícia que foi classificada como falsa, devido ao fator de ocorrências das palavras encontradas nas notícias levantadas pelo *crawler* ser alto.

Utilizando pesquisas realizadas por usuários armazenadas no banco de dados, foram levantadas no gráfico da Figura 6, as porcentagens de notícias pesquisadas onde se tratam de notícias de conteúdo imediato nos meses de março a novembro de 2021. Nenhuma notícia apresentada foi classificada com possível chance de se tratar de *fake news*.

Figura 6 – Notícias de conteúdo imediato



Fonte: Criado pelo autor

Como foi possível observar nos resultados encontrados, algumas notícias trazem uma porcentagem interessante para o usuário, onde o mesmo possa ao menos ter uma breve concepção da notícia e com indicativos, onde uma notícia pode ser verdadeira ou falsa. As imprecisões da porcentagem ainda ocorrem quando tratamos uma notícia falsa, devido ao alto índice de ocorrência de palavras encontradas textualmente. O mesmo pode acontecer para uma

notícia verdadeira, onde ao ser pesquisada, a falta de ocorrências da sentença inserida pelo usuário, pode não acarretar em um resultado alto da porcentagem.

Em todos os casos em que foi tratado uma notícia falsa ao decorrer do projeto, encontrou-se alguma palavra que indicava ou associava que aquela notícia poderia se tratar de uma *fake news*, dessa forma apontava para o usuário que independente da porcentagem apresentada poderia se tratar de uma notícia falsa.

6. Considerações finais

Podemos concluir que a partir desta aplicação desenvolvida, a mesclagem de algoritmos, até mesmo sem usar um aprendizado de máquina e buscando trazer uma forma mais humana ao realizar as análises textuais, podem trazer resultados satisfatórios. Ao explorar todos os caminhos possíveis de processamento de linguagem natural atrelado com técnicas apuradas para obter a ocorrência de palavras, resultados mais concisos poderão ser obtidos.

Ao longo do projeto, pensando em utilizá-lo no futuro com as notícias atualizadas, foi descartado a utilização de um *dataset,* pois seria tratado as notícias mais recentes e atualizadas diretamente pelo *crawler,* já que os mecanismos de buscas trazem os resultados mais precisos e recentes da *web.*

Por meio de informações obtidas e tratadas pelos mecanismos de buscas, podemos oferecer para o usuário uma breve concepção caso as notícias que foram pesquisadas se tratem de notícias falsas ou não. Portanto, é um mecanismo que serve de apoio para o usuário identificar os fatos, visando contribuir com que o ecossistema de notícias mantenha sua autenticidade.

A partir do que foi atingido neste trabalho há melhorias e desenvolvimentos a serem realizados para futuras versões, como: abordar todas as formas possíveis de linguagem natural e a utilização de tecnologias computacionais como *machine learning* para que o sistema consiga obter precisões cada vez maiores.

REFERÊNCIAS

ALLCOTT, H., & GENTZKOW, M. **Social media and fake news in the 2016 election**. 2017. 102 f. Journal of Economic Perspectives. Disponível em: https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.31.2.211. Acesso em: 16 jun 2021.

BASHEER, I. A.; HAJMEER, M. **Artificial neural networks: fundamentals, computing, design, and application.** 2020. 29f.

BROWNLEE, J.; **What Are Word Embeddings for Text?**. Machine Learning Mastery. 11 de out. de 2017. Disponível em: https://machinelearningmastery.com/what-are-word-embeddings. Acesso em: 25 jan 2021.

BOBERG, T. Q. L. F. S.; SCHATTO-ECKRODT, R. R. **Fake News.** 2019. 6f. Disponível em: https://www.researchgate.net/publication/332749986. Acesso em: 17 jan 2021.

COSTA, L. G. **Classificação de fake news utilizando algoritmos de aprendizado de máquina e aprendizado profundo**. 2019. 82 f. Trabalho de Conclusão de Curso (Bacharel em Ciência da Computação) - Universidade Federal de Roraima, Boa Vista, 2019.

DELMAZO, S. C. **FAKE NEWS NAS REDES SOCIAIS ONLINE: PROPAGAÇÃO E REAÇÕES À DESINFORMAÇÃO EM BUSCA DE CLIQUES**. 2018. 15 f. Artigo Científico - Universidade Nova de Lisboa, Lisboa, 2018. Disponível em: https://impactum-journals.uc.pt/mj/article/view/2183-5462_32_11/4561. Acesso em: 14 jan 2021.

FERREIRA, P. A.; **O avanço da tecnologia e as transformações na sociedade.** Notícias Portal Da Industria. 11 de out. de 2017. Disponível em: https://noticias.portaldaindustria.com.br/artigos/paulo-afonso-ferreira/o-avanco-da-tecnologia-e-as-transformacoes-na-sociedade/ Acesso em: 05 set 2021.

GADE, P.; RAVIOLA, E. **Integration of News and News of Integration: A Structural Perspective on News Media Changes.** 2015. 26f.

Galbraith, J. R. **Competing with flexible lateral organizations**. Boston, MA: Addison-Wesley. 1994.

GRUPPI, M.; HORNE, B. D.; ADALI, S. **An Exploration of Unreliable News Classification in Brazil and The U.S.** 2018. 5 f. Rensselaer Polytechnic Institute, Troy, New York, USA, 2019. Disponível em: https://arxiv.org/pdf/1806.02875.pdf. Acesso em: 10 jan 2021.

HOROWITZ, B. T. **Can AI Stop People From Believing Fake News?** Machine learning algorithms provide a way to detect misinformation based on writing style and how articles are shared. IEEE Spectrum. 15 de mar. de 2021. Disponível em: https://spectrum.ieee.org/ai-misinformation-fake-news. Acesso em: 12 ago 2021.

HUI-CHANG, W.; SHU-HUA, R.; QI-JIE, T. **The Implementation of a Web Crawler URL Filter Algorithm Based on Caching**. 2009. 4f.

KARIMI, H.; ROY, P. C.; SABA-SADIYA, S.; TANG, J. **Multi-Source Multi-Class Fake News Detection**. 2018. 12f. Disponível em: https://aclanthology.org/C18-1131.pdf. Acesso em: 13 jan 2021.

MANJOO, F. **How the Internet Is Loosening Our Grip on the Truth**. The New York Times. 02 de nov. de 2016. Disponível em:

https://www.nytimes.com/2016/11/03/technology/how-the-internet-is-loosening-our-grip-on-the-truth.html?smid=pl-share. Acesso em: 11 ago 2021.

MONTEIRO, R. O.; NOGUEIRA, R. R. **Projeto de um Sistema Web a Classificação de Fake News** In: CONGRESSO LATINO-AMERICANO DE SOFTWARE LIVRE E TECNOLOGIAS ABERTAS (LATINOWARE), 16., 2019, Foz do Iguaçu. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2019. p. 120-123. Disponível em: https://sol.sbc.org.br/index.php/latinoware/article/view/10343. Acesso em: 12 jun 2021.

PÉREZ-ROSAS, V.; KLEINBERG, B.; LEFEVRE, A.; MIHALCEA, R. **Automatic Detection of Fake News.** 2017. 11f. Disponível em: https://aclanthology.org/C18-1287.pdf Acesso em: 05 abr 2021.

SHU, K.; SILVA, A.; TANG, J.; LIU, H. **Fake News Detection on Social Media: A Data Mining Perspective**. 2017. 15 f. Disponível em: https://www.researchgate.net/publication/318981549_Fake_News_Detection_on_Social_Media_A_Data_Mining_Perspective. Acesso em: 02 nov 2020.

SILVA, T. S. **O MONOPÓLIO DA VERDADE NA ERA DAS FAKE NEWS**. 2019. 19 f. Universidad Autonoma Latinoamericana, 2019. Disponível em: http://www.redalyc.org/articulo.oa?id=585762914004. Acesso em: 10 jan 2021.

SIVEK, S. C. **Both facts and feelings: Emotion and news literacy**. 2018. 16 f. Journal of Media Literacy Education. Disponível em: https://digitalcommons.uri.edu/cgi/viewcontent.cgi?article=1355&context=jmle. Acesso em: 15 jun 2021.

WARDLE, Claire. **Fake news. It's complicated.** First Draft, 16 de fev. de 2017. Disponível em: https://firstdraftnews.org/articles/fake-news-complicated/. Acesso em: 01 set 2021.