DOI: 10.1590/s0103-4014.2021.35101.009

A Inteligência Artificial e os desafios da Ciência Forense Digital no século XXI

Rafael Padilha, [I] *Antônio Theóphilo,* [II] *Fernanda A. Andaló,* [III] *Didier A. Vega-Oliveros,* [IV] *João P. Cardenuto,* [V] *Gabriel Bertocco,* [VI] *José Nascimento,* [VII] *Jing Yang,* [VIII] *e Anderson Rocha* [IX]

Introdução

A ciência forense digital (CFD) é o ramo da Ciência Forense que trata da análise e investigação de conteúdos associados a dispositivos digitais (Casey, 2011), desde equipamentos de grande porte, como servidores e roteadores de empresas globais, até dispositivos móveis, como celulares e *wearables*.

Juntamente ao desenvolvimento da computação pessoal desde a década de 1970, a CFD tem enfrentado diversos desafios, pela própria evolução das tecnologias digitais e do seu material alvo, e.g., mídias, dados e *posts*. Contudo, nenhum desafio foi tão intenso quanto o surgimento das mídias sociais e os avanços da Inteligência Artificial (IA) nos últimos anos.

Até pouco tempo atrás, as análises feitas por um especialista forense se baseavam predominantemente em provas adquiridas durante investigações, relatos e conteúdo de TV e jornais. Atualmente, o cenário é diferente. Com a popularização de dispositivos móveis, qualquer pessoa pode registrar informações relacionadas a um acontecimento – seja um protesto, seja um atentado, seja um evento de entretenimento. Da mesma forma, pessoas em qualquer parte do mundo podem interagir e propagar essas informações.

Uma análise forense que anteriormente se limitava a um conjunto reduzido de fontes hoje em dia precisa lidar com a infinitude de dados compartilhados nas redes sociais. Eventos de alcance regional produzem milhares de *posts* e imagens, como a recente explosão em Beirute, que gerou mais de 150 milhões de *tweets* e mais de 17 mil imagens e vídeos. A situação é ainda mais proeminente em eventos de escala global, como a recente pandemia provocada pelo coronavírus. Diversos conjuntos de dados já estão disponíveis, alguns contando com 920 milhões de documentos (Chen; Lerman; Ferrara, 2020), permitindo inclusive a análise da pandemia sob o aspecto da desinformação (Cinelli et al., 2020).

Tal volume de dados gera inúmeros desafios nas etapas do processo forense (Padilha et al., 2020). Uma vez que os dados referentes a um evento – como imagens, vídeos e texto – tenham sido coletados, a primeira etapa consiste em filtrar as informações relevantes. Parte do conteúdo compartilhado sobre um evento é composta não só pelas informações relevantes, mas também por *memes*, montagens e opiniões de usuários que nem sempre retratam a realidade. Filtrar as informações relevantes, a partir de um conjunto praticamente infinito de dados, é uma etapa essencial para garantir a confiança de análises posteriores.

A etapa seguinte procura organizar semanticamente os dados relevantes. A natureza da organização pode variar segundo as características do evento, geralmente envolvendo sincronização cronológica e espacial dos dados. Esse processo é crucial, pois traça relações entre as informações, fornecendo uma visão geral do evento, e permite que o analista forense foque em determinados atores ou momentos que sejam pertinentes.

Na última etapa, o papel do perito é reconstruir a narrativa do evento a partir dos dados organizados, de modo a compreender o que aconteceu, quais foram os agentes que participaram, e sob quais condições o evento se desencadeou. Para isso, é necessário relacionar as informações disponíveis entre si, tarefa que é dificultada pelo volume de dados que supera em muito a capacidade humana de análise, mesmo depois de filtrados.

Além da massiva disponibilidade de dados digitais, a evolução dos algoritmos de IA e o desenvolvimento de plataformas de computação paralela mais acessíveis desempenham um papel importante. O mais recente e surpreendente avanço são as *Deep Fakes* (Chesney; Citron, 2019), modelos de IA baseados em redes de aprendizado profundo (*Deep Learning*) capazes de gerar áudios e vídeos falsos de pessoas reais com alto grau de realismo. Essa nova e assustadora realidade motivou a agência de defesa americana Darpa a lançar o programa SemaFor[1] para atualizar os esforços do programa anterior MediFor[2] que abordava apenas os cenários de manipulação tradicionais, como a inclusão pontual de elementos artificiais em textos, áudios ou imagens.[3] A União Europeia possui esforços similares através do Projeto Reveal,[4] com foco no conteúdo compartilhado em mídias sociais.

Até mesmo desafios recentes e aparentemente desconexos com a CFD, como a detecção de notícias falsas (as chamadas *Fake News*), podem ser vistos como problemas dessa área (Ferreira et al., 2019). Quando milhares de notícias são publicadas sobre um evento, parte delas expõe uma visão falsa ou levemente manipulada dos fatos, geralmente na tentativa de reformular o que ocorreu ou desviar o foco. Uma análise *post-factum* pode responder a diversas perguntas: em que fatos a notícia suspeita se baseia, quem a escreveu, que outros textos estão relacionados. Em Omezi e Jahankhani (2020), os autores analisam o processo atual de checagem de fatos (*fact-checking*) e como ele pode ser alinhado com procedimentos forenses já consagrados, como o ACPO Forensic Guide-

lines (Sammes; Jenkinson, 2007). Outro trabalho (Middleton; Papadopoulos; Kompatsiaris, 2018) propõe a integração e o uso de ferramentas de análise forense digital por jornalistas, como um processo rápido e de baixo custo para a checagem de notícias.

Essa mesma realidade que desafia a CFD também indica o caminho para enfrentá-la. Os problemas mencionados anteriormente são abordados pela comunidade científica com métodos modernos de IA. No que tange à coleta e filtragem de dados publicados *online*, diversas técnicas de Processamento de Linguagem Natural (PLN) e visão computacional são imprescindíveis para processar as informações em tempo hábil. Durante a análise dos dados, essas técnicas são ainda mais importantes, uma vez que tiram proveito da imensa quantidade de dados disponíveis para encontrar padrões capazes de complementar a análise forense.

Entretanto, as soluções ainda apresentam questões-chave que precisam ser tratadas. Uma delas é interpretabilidade dos métodos, ou XAI (*eXplainable Artificial Intelligence*) (Gunning et al., 2019). À medida que os modelos se tornam mais poderosos e descritivos, perdemos a capacidade de interpretar seu fluxo de decisão. Essa limitação no entendimento desses modelos "caixa-preta" pode restringir o uso em problemas reais. Na CFD é necessário identificar o motivo pelo qual um modelo tomou uma decisão, especialmente se essa análise for utilizada em investigações e processos judiciais.

Outro desafio para a aplicação de técnicas de IA no contexto da CFD é a mitigação de vieses. Modelos de aprendizado de máquina aprendem complexas funções de decisão a partir de dados reais. Por serem um recorte de nossa sociedade, muitas vezes esses dados contêm vieses que não devem ser reproduzidos. Exemplos recentes incluem a tendência de sistemas de tradução automática de texto, como o Google Tradutor, em reproduzir vieses de gênero (Prates; Avelar; Lamb, 2019). Palavras como *doctor* e *engineer* eram comumente associadas a artigos, pronomes e sufixos masculinos ("o médico", "o engenheiro"), enquanto *nurse* e *baker* eram relacionadas a termos femininos ("a enfermeira", "a confeiteira"). Identificar e mitigar esses vieses é um problema atual não só para a CFD, mas para toda a comunidade de IA.

Este artigo tem como objetivo apresentar e discutir como o advento das mídias sociais e os avanços da IA impactaram a CFD nos últimos anos. Sobre o pano de fundo da análise forense de eventos – seja ele no mundo físico, seja no digital – abordamos as etapas e os métodos utilizados para coletar, filtrar, organizar semanticamente e analisar os dados disponíveis. Na segunda seção, discutimos o protocolo de análise forense de eventos, detalhando cada etapa e seus objetivos. Em seguida, examinamos sua aplicação e os principais desafios que surgem ao analisar eventos no mundo físico (terceira seção) e virtual (quarta seção), abordando trabalhos da literatura relevantes para o problema. Apesar da divisão aparente entres os dois mundos, discutimos na quinta seção como a

análise conjunta de ambos é fundamental para a reconstrução e compreensão de um evento. Finalmente, detalhamos na sexta seção alguns dos desafios do uso de métodos de IA em contextos sensíveis, como a CFD.

Ciência Forense na era das mídias sociais e da Inteligência Artificial

O "Princípio da Transferência", cunhado por Edmond Locard no início do século XX, de que "todo contato deixa rastros" serviu de base para o desenvolvimento da Ciência Forense moderna (Chisum; Turvey, 2000). Ele estabelece que toda interação entre dois agentes – pessoas, objetos e lugares – deixa algum tipo de evidência que pode ser encontrada e analisada para se compreender o contexto que a envolve. Apesar de Locard se referir a rastros deixados no mundo físico, a mesma ideia vale para o mundo virtual. Onde anteriormente havia pistas como impressões digitais, pegadas ou sangue, hoje também entram em cena fotos, clipes de áudio/vídeo, *posts* em redes sociais, transações bancárias, registros de localização (GPS), entre outros. Essas novas fontes de informações implicaram uma reformulação de várias áreas forenses, como a Ciência Forense Social (Mcfarland; Lewis; Goldberg, 2016) e a Medicina Forense (Lefèvre, 2018). Como previsto por (Pollitt, 2010), a análise forense deixou de ser um processo linear de recuperação de evidências e se tornou um processo integrado de organização de informações que exige de seus especialistas abordagens multidisciplinares para se extrair relações do enorme volume de dados.

De forma semelhante, a CFD também precisou ser reformulada para se encaixar nessa nova realidade. Até pouco tempo atrás, eventos que ocorriam no mundo físico eram analisados com procedimentos tradicionais da Ciência Forense (e.g., entrevista de testemunhas, coleta de impressões digitais, entre outros). Contudo, dada a onipresença das mídias sociais em nossas vidas, é muito provável que fotos, vídeos e relatos de um evento tenham sido disponibilizados nessas mídias. Esse fluxo de informações – que surge tanto durante o acontecimento como também posteriormente, à medida que as informações são disseminadas entre usuários e mídias – superam a capacidade humana de análise. Os métodos de IA se mostram como ferramentas necessárias para superar essas limitações, buscando reduzir o escopo em um conjunto de dados relevantes e organizá-los de modo que a experiência e o conhecimento do perito possam ser mais bem aproveitados (Pollitt, 2010).

Um exemplo do uso dessas ferramentas ocorreu durante as investigações sobre a invasão do Capitólio nos Estados Unidos (janeiro, 2021). Incentivados por um discurso do ex-presidente Donald Trump, milhares de seus apoiadores invadiram o prédio do governo e entraram em confronto com a polícia. Além de ser noticiado pela mídia tradicional, o ato foi documentado por meio de imagens e vídeos compartilhados ao vivo pelos próprios invasores. Apesar de os vídeos capturarem os invasores à medida que adentravam as salas do Capitólio, seria necessário um esforço considerável para recuperar suas identidades. Dessa forma,

investigadores utilizaram algoritmos de reconhecimento facial para identificar possíveis suspeitos, associando os vídeos produzidos com fotos disponíveis na internet.[5] Assim como, em um primeiro momento, os dados auxiliaram na busca dos suspeitos, uma análise mais profunda também pode ajudar os investigadores a responder diversas questões acerca da invasão. Por exemplo, como os manifestantes se organizaram, que tipos de armas portavam, quais notícias falsas foram compartilhadas sobre o evento, entre outras.

Para que análises aprofundadas possam ser realizadas, no entanto, é necessário assegurar que as informações necessárias para entender o que aconteceu sejam recuperadas e organizadas, de modo que o perito chegue a conclusões verdadeiras. Dessa forma, é essencial estabelecer um rigoroso protocolo de análise, como o mostrado na Figura 1, cujas etapas são:

Coleta de dados: para um evento de interesse em particular, é realizada a aquisição de dados disponíveis publicamente por meio de fontes variadas, como mídias sociais, veículos tradicionais de jornalismo e sistemas de vigilância. Nessa etapa, busca-se pela diversidade de modalidades e fontes que possam fornecer às etapas seguintes pontos de vista complementares sobre o evento.

Sanitização e filtragem de relevância: apesar da enorme disponibilidade, nem todo dado coletado é relevante para uma análise forense. Entre o conteúdo adquirido, há todo o tipo de informação irrelevante, como *memes*, fotos manipuladas, opiniões textuais e informações de eventos passados. Além disso, há uma enorme redundância nos dados, muitas vezes com sutis modificações no conteúdo original, aumentando o custo computacional em processá-los. O propósito dessa etapa é reduzir os dados a um subconjunto relevante para a análise, a partir de uma descrição feita por um especialista forense do que é importante.

Organização semântica: quanto mais um dado é difundido pela rede, mais difícil se torna rastrear o seu ponto de origem, as modificações que sofreu durante os compartilhamentos e a relação semântica com outros dados. Nessa etapa, esses laços são recuperados, buscando sincronizar as informações cronológica e geograficamente em relação ao evento, agrupando dados semanticamente semelhantes, e elencando potenciais atores, objetos, lugares e momentos de destaque dentro do acontecimento.

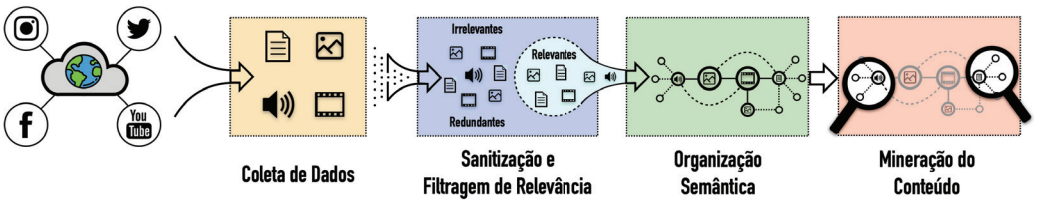
Mineração do conteúdo: a partir de uma organização coerente, analisa-se quais conhecimentos podem ser extraídos dos dados. As análises dessa etapa variam consideravelmente, dependendo da natureza do evento alvo. Por exemplo, no caso de um atentado terrorista, pode ser importante detectar pessoas e objetos de interesse, acompanhando suas trajetórias no espaço e no tempo pelo evento.

Eventos do mundo físico são, geralmente, registrados em fotos e vídeos, que são prontamente compartilhados nas redes, junto com relatos do que aconteceu. Esses dados requerem uma atenção especial durante sua organização, para que sejam sincronizados espacial e cronologicamente em relação ao evento e possam auxiliar o entendimento do que realmente ocorreu. Por sua vez, eventos

no mundo virtual nascem geralmente de *posts* e discussões em mídias sociais. As análises desse tipo de evento visam compreender como ele se originou e como repercutiu no mundo virtual. À medida que as informações se propagam pelas redes, elas são modificadas pelas interações com os usuários, que incorporam a sua visão pessoal a cada compartilhamento. Nesse contexto, buscamos verificar como se deram essas interações, traçar relações entre comentários que surgiram a partir do evento, identificar os principais atores, possíveis focos de desinformação e também qual a percepção sobre o evento em diferentes *bolhas* (grupos isolados de usuários com a mesma orientação política) pela rede.

Apesar de os mundos físico e virtual, na teoria, estarem separados pelas telas de dispositivos eletrônicos, na prática, todo evento no mundo real produz sua contraparte digital, na forma de notícias compartilhadas e discussões entre usuários de redes sociais. De forma semelhante, eventos originários do mundo virtual podem ecoar no mundo real, a partir de boatos, notícias falsas, entre outros. Por exemplo, em 2014, uma mulher foi morta por vizinhos em Guarujá (SP) por causa de boatos em mídias sociais.[6] Após relatos mentirosos de sequestros que aconteciam na região, vizinhos consideraram que a mulher se parecia com um retrato falado divulgado na rede social e a lincharam. Nesses casos, a análise conjunta de ambos os mundos é essencial para recuperar informações complementares para a compreensão total do acontecimento.

Impulsionada pela disponibilidade massiva de dados, a comunidade científica de CFD e IA exploraram as particularidades desse novo cenário de diferentes formas. Nas próximas seções, apresentamos e discutimos os desafios e oportunidades existentes nas análises de mundo físico e virtual.



Fonte: Elaboração própria.

Figura 1 – O protocolo de análise forense busca guiar a análise no contexto atual das mídias sociais. Algoritmos e modelos de IA têm o papel de auxiliar o perito, automatizando processos em todas as etapas apresentadas.

Análise do mundo físico

Nesta era digital, praticamente tudo o que ocorre no mundo real vira *bit* e é compartilhado. Especialmente quando a atenção global se volta para eventos de grande escala, o volume de dados compartilhados é surpreendente. Um exemplo é o atentado à maratona de Boston, quando duas bombas explodiram na linha de chegada, em 2013. O Twitter foi inundado por mais de 700 mil menções ao ataque em menos de duas horas, incluindo imagens e vídeos capturados no evento.[7]

É possível supor que esses dados sejam uma rica fonte de informações para entendermos melhor o mundo físico, permitindo a checagem de fatos e investigações. Para isso, a análise de conteúdo possibilita a estruturação dos itens compartilhados e posterior ligação ao evento físico. Os passos dessa análise (Figura 1) são realizáveis por meio de técnicas de IA, já que a complexidade excede a capacidade humana, especialmente em curto prazo.

O fogo que destruiu a Grenfell Tower, em Londres, é um exemplo de evento físico cuja análise, a partir dos itens digitais, levou a um melhor entendimento do ocorrido. Em 2017, o incêndio que devastou o prédio residencial foi compartilhado ao vivo por diversos transeuntes e pela mídia. Um estudo[8] aplicou uma variação das etapas mencionadas para organizar os itens digitais em um mesmo sistema de coordenadas, permitindo uma visão geral da tragédia.

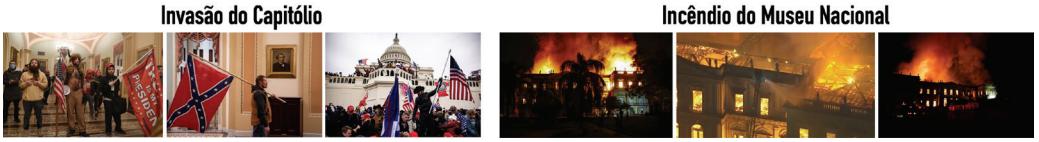
Mas mesmo que esse tipo de análise já tenha sido realizado com sucesso, ainda restam as perguntas de como pode ser aplicado a outros tipos de evento e quais técnicas podem ser utilizadas. Considerando dois eventos recentes e completamente distintos – o incêndio do Museu Nacional no Rio de Janeiro e a invasão do Capitólio americano – vamos analisar as etapas, considerações específicas e possíveis técnicas.

Coleta e filtragem de dados

Em 2018, o Museu Nacional, edifício histórico no Rio de Janeiro, foi extremamente danificado por um incêndio de grandes proporções, que levou consigo quase todo o seu acervo. As imagens e os vídeos do evento são focados no prédio e capturados durante aproximadamente sete horas, mostrando a evolução do fogo e suas consequências na fachada, além da movimentação de pessoas. Já na recente invasão ao Capitólio americano, apesar de muitas informações estarem focadas no prédio em si, grande parte do evento se desenrola nos arredores e dentro do prédio, tornando a análise ainda mais desafiadora. O objetivo do entendimento dos dois eventos também é distinto. No primeiro caso, procura-se compreender a dinâmica da destruição do prédio pelo fogo e como foi combatido. No segundo, procura-se identificar suspeitos e falhas de segurança.

Apesar da discrepância entre os dois eventos, as técnicas de coleta e filtragem de dados são agnósticas às diferenças. O objetivo principal é coletar e filtrar itens que são *representativos*, i.e., relevantes ao evento sendo analisado, rejeitando os *não-representativos* (Figura 2).

Nesse sentido, redes sociais são uma boa fonte de conteúdo, contendo materiais tanto de veículos de mídia quanto de pessoas que registraram fotos e vídeos por meio de seus dispositivos móveis. A coleta desses dados pode ocorrer por meio de API (interfaces programáveis de aplicações) fornecidas pelas próprias redes ou por meio de raspagem (*crawling*). Em termos gerais, a segunda abordagem apresenta melhores resultados, pois é independente de restrições impostas pelas redes, como a vazão da coleta (número de itens coletados em um determinado período de tempo) (Hernandez-Suarez et al., 2018).



(a) Imagens representativas capturam o local de interesse do evento durante ou logo após o acontecimento.

(b) Imagens não representativas podem conter memes, montagens, desenhos, fotos de outros locais ou até mesmo do local de interesse, porém fora da duração do evento.

Fonte: Elaboração própria.

Figura 2 – Exemplos de imagens representativas e não-representativas para os dois eventos estudados.

Independentemente da abordagem da coleta, apenas uma fração dos dados coletados é representativa, pois muitos itens são cópias, montagens ou não são relacionados (Gupta; Lamba; Kumaraguru, 2013). Para a detecção e posterior remoção de duplicatas ou duplicatas próximas, a aplicação de *hashing* de imagens é uma técnica eficaz. Mas, dado que a remoção de imagens não representativas é um problema fundamentalmente de classificação, a utilização de redes neurais convolucionais (Convolutional Neural Network - CNN) é mais apropriada.

O *hashing* de imagens para detecção de imagens que são aproximadamente cópias pode ser feito por meio de Perceptual Hash (Schneider; Chang, 1996) ou Wavelets Hash (Venkatesan et al., 2000). Tendo sido identificado um par de imagens como duplicatas próximas, definir qual deve ser mantida no conjunto de imagens relevantes é um problema em aberto (Padilha et al., 2021). Já as CNN podem ser aplicadas de variadas formas. Dois exemplos encontrados na literatura são: combinação de diferentes redes, sendo cada uma especializada em identificar uma característica diferente (como pessoas, lugares, ou objetos), e a filtragem baseada na combinação desses aspectos (Rodrigues et al., 2019); e filtragem em três passos: uma rede filtra imagens que são inerentemente irrelevantes (e.g., desenhos), depois imagens duplicadas são removidas por Perceptual Hashing e, por fim, as imagens restantes tem sua relevância avaliada por outra rede (Nguyen et al., 2017).

Esses algoritmos normalmente necessitam que alguns exemplos de imagens representativas (*sementes*) sejam fornecidos. Porém, a rotulagem manual não é uma tarefa simples. Por exemplo, no caso de o evento ser um protesto de rua, uma imagem de outro protesto ocorrido no mesmo lugar poderia ser atribuída como relevante. Por essa razão, a experiência do perito é fundamental, tanto na forma de análise de alguns dados isoladamente, como fornecendo as sementes.

Fonte: Elaboração própria.

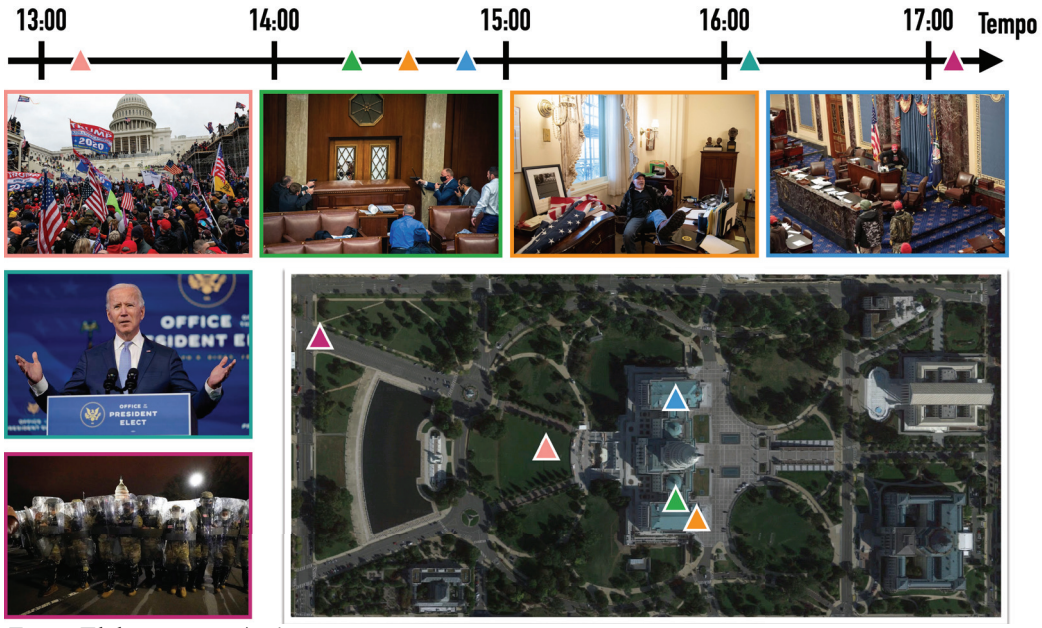
Figura 3 – Divisão do incêndio no Museu Nacional em intervalos fixos de uma hora ou em subeventos. A organização por intervalos é flexível e exige pouco conhecimento sobre o evento. No segundo caso, é necessário entender o evento para definir os subeventos, mas a abordagem facilita a organização das imagens devido às características mais marcantes entre os subeventos.

Organização semântica

Essa etapa tem como objetivo estruturar os dados já coletados e filtrados, a fim de encontrar uma ligação entre eles, bem como com o evento sendo analisado. A ideia é construir uma visão geral do evento e poder recontar a história a partir do que foi compartilhado previamente.

Um exemplo de estruturação relevante aos dois eventos estudados é a organização temporal dos itens coletados. Porém, essa tarefa precisa ser realizada levando-se em consideração as características de cada evento.

Para o incêndio do Museu Nacional, pode-se considerar o problema de organização temporal sob duas ópticas (Figura 3). A primeira é classificar as imagens em intervalos de tempo fixos que se estendem pela duração total do evento. Sabe-se que o incêndio teve início por volta das 19h30, apenas sendo controlado às 2h da madrugada do dia seguinte. Após esse período, os bombeiros continuaram no local combatendo outros focos de fogo. Há muitas imagens capturadas no dia seguinte ao incêndio, mostrando a estrutura danificada do prédio. Todo esse período de tempo pode ser dividido em intervalos de, por exemplo, uma hora, e métodos de IA serem utilizados para classificar as diferentes imagens nesses intervalos. Um agravante é que a maioria das imagens é noturna, dificultando a análise visual pela diminuição das dicas temporais, como mudança de luminosidade. Não obstante, métodos de IA podem se guiar por outras dicas, como a evolução do fogo ou ainda a degradação da fachada do prédio ao longo do tempo.



Fonte: Elaboração própria.

Figura 4 – Exemplo de visualização para a invasão ao Capitólio, com as organizações espacial e temporal. O objetivo é fornecer ao perito uma visão geral do evento que permita análises posteriores.

A segunda opção é a divisão do evento em subeventos significativos: do início do incêndio até a chegada do Corpo de Bombeiros, o aumento do dano na fachada pelo fogo, desabamento de parte do teto, controle do incêndio, e fogo completamente extinto. Método similar foi utilizado na organização do incêndio da Catedral de Notre-Dame (Padilha; Andaló; Rocha, 2020), no qual o evento foi subdividido em subeventos marcados por mudanças na estrutura da catedral e um classificador foi treinado para classificar imagens nesses subeventos.

Já a invasão do Capitólio americano ocorre em diferentes cenários: tem início nos arredores do prédio, logo após ocorre a invasão e há subeventos em diferentes partes internas. Nesse sentido, pode-se considerar a mesma técnica usada na análise da explosão na Maratona de Boston (Pinheiro et al., 2019; Lameri et al., 2014). O objetivo é sincronizar temporalmente vídeos do mesmo evento capturados por diferentes câmeras. O aspecto forense importante é poder estimar quais quadros de quais vídeos contêm informação sobre um subevento importante. Assim, é possível observar o mesmo subevento sob diferentes pontos de vista. O método em Pinheiro et al. (2019) extrai assinaturas características de cada vídeo, por meio de uma CNN, e compara essas assinaturas a fim de detectar sequências que se sobrepõem.

Um outro exemplo de estruturação relevante é a organização espacial. No evento do Capitólio, pode ser interessante para um investigador saber, por exemplo, qual fachada ou parte interna do prédio está em foco a cada intervalo

de tempo ou subevento. Semelhantemente a Padilha et al. (2021), é possível treinar um classificador para identificar o cenário ou fachada sendo representado em cada imagem. Com as organizações temporal e espacial realizadas, pode-se representar todo o evento em um mesmo sistema de coordenadas, ou visualização. Por exemplo, os itens podem ser apresentados em um mapa dinâmico, com a localização de cada item ao longo do tempo (Figura 4).

Mineração de conteúdo

Ao lidar com problemas forenses, uma pergunta importante é "quem participou do evento?". Na invasão do Capitólio nos Estados Unidos, por exemplo, a principal investigação foi a busca pelos suspeitos que apareceram nas imagens e vídeos compartilhados.[9] Nesse sentido, podem-se utilizar os itens já organizados semanticamente para encontrar pessoas, objetos e lugares envolvidos e, ainda, propor alguns candidatos para investigação posterior.

Uma possível técnica para responder às perguntas forenses é a de reidentificação de pessoas por meio de agrupamento (Zhai et al., 2020; Song et al., 2020), com a qual se busca reconhecer a mesma pessoa em diferentes momentos e locais do evento, de forma a criar uma análise do seu comportamento ao longo tempo.

Para a reidentificação, é necessário primeiramente detectar, em cada imagem do evento, o maior número de pessoas possível e aplicar um modelo de extração de características para obter uma representação vetorial para cada indivíduo. No entanto, os itens compartilhados estão sujeitos a diferentes condições (como iluminação, posição, resolução), o que pode gerar uma representação significativamente diferente para a mesma pessoa. Durante a invasão ao Capitólio, por exemplo, as pessoas estavam, inicialmente, ao redor do prédio e, portanto, poderiam estar sendo filmadas por câmeras de seguranças externas. Após a invasão, passam a ser monitoradas pelas câmeras internas sob condições de iluminação e plano de fundo diferentes. Para ser possível casar a representação vetorial da mesma pessoa fora do Capitólio com suas respectivas imagens internas, é primordial a utilização de modelos com alto grau de descrição semântica. Nesse sentido, como em outras etapas já mencionadas, CNN podem ser aplicadas.

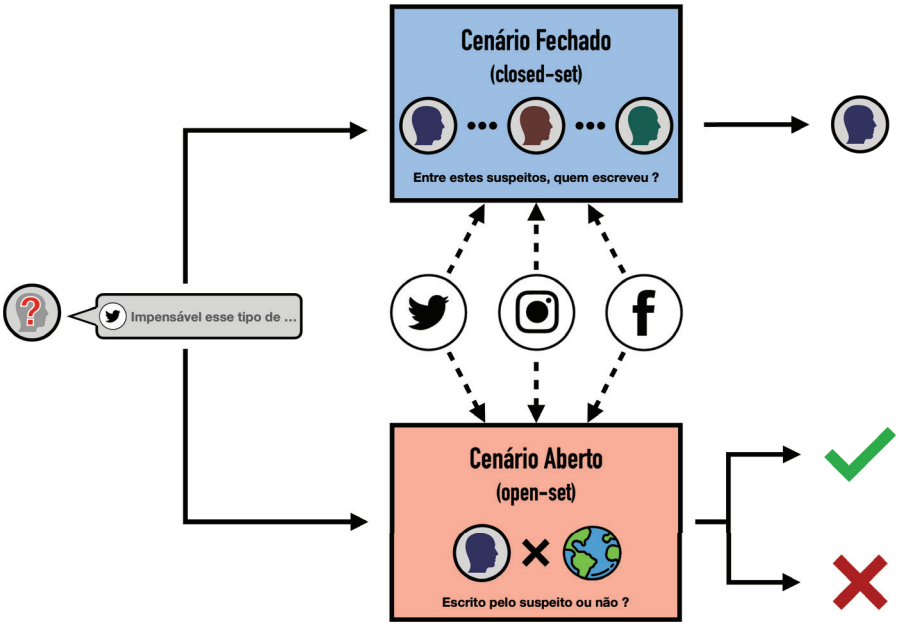
É desejável que o treinamento das CNN seja feito sem a exigência de que todas as pessoas sejam previamente rotuladas. Nesse sentido, as técnicas de transferência de conhecimento e as de adaptação entre domínios (*Domain Adaptation*) são uma alternativa promissora. Mas como o conhecimento do modelo provém de uma base de dados potencialmente distinta do evento estudado, é natural que muitos grupos ruidosos sejam detectados (Yang et al., 2020; Ester et al., 1996). Assim, é necessário realizar um processo de filtragem de grupos, mantendo apenas aqueles de maior confiança, e de extração de novas representações vetoriais para esses grupos. Esse processo de adaptação é aplicado de forma iterativa, gerando, ao final, um modelo melhor em agrupar pessoas no evento de interesse. Este mesmo processo pode ser estendido para objetos.

Uma vez obtidos os grupos de interesse, o evento pode ser estudado por meio do entendimento conjunto desses diversos grupos. No caso da invasão do Capitólio, podem-se identificar grupos de pessoas próximas entre si ao longo da invasão e obter informações de objetos que possam ter sido utilizados, como veículos, mochilas, ou armas. No incêndio do Museu Nacional, pode-se entender a dinâmica do combate ao incêndio a partir da identificação das pessoas ao redor do prédio.

Análise do mundo virtual

O desenvolvimento tecnológico e o crescimento das mídias sociais ocorridos nos últimos anos promovem, cada vez mais, o surgimento de cenários que envolvem apenas o mundo virtual e que também demandam tratamento da CFD. Crimes como racismo, misoginia, *bullying*, *phishing*, plágio, desinformação e manipulação da opinião pública em larga escala[10] ocorrem frequentemente no ambiente virtual, fornecendo vestígios digitais que precisam ser utilizados para elucidá-los.

Esta seção apresenta alguns exemplos de cenários do mundo virtual (frequentemente envolvendo a presença de mídias sociais) onde a CFD tem contribuído para endereçar os problemas e crimes mencionados anteriormente, quase sempre lançando mão de modernas técnicas de IA.



Fonte: Elaboração própria.

Figura 5 – Diferentes cenários de atribuição de autoria em mídias sociais. No cenário *fechado*, uma mensagem precisa ser atribuída a um dos suspeitos. No cenário *aberto*, temos um único suspeito e precisa-se decidir se a mensagem foi escrita por ele ou por outra pessoa desconhecida. Em ambos os casos, modelos de autoria baseados em IA são treinados usando dados de mídias sociais.

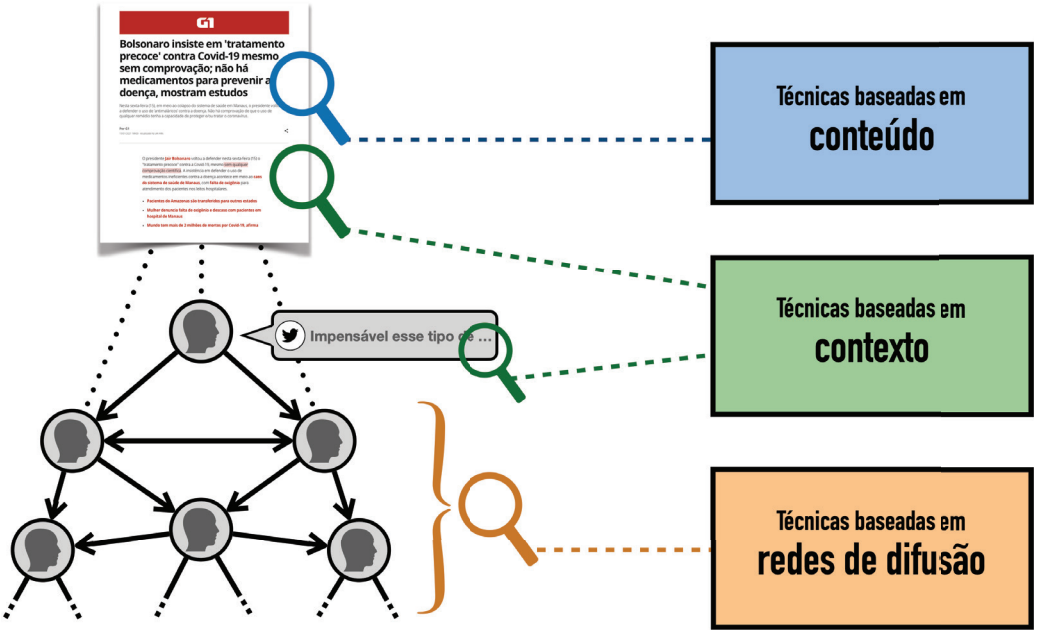
Atribuição de autoria em mídias sociais

Atualmente, grande parte da informação que consumimos diariamente advém das mídias sociais, tanto proveniente de perfis associados aos grandes veículos de comunicação como também de perfis menos conhecidos. Essa nova realidade, apesar de mais democrática, abre espaço para o fortalecimento de muitos dos crimes mencionados no início desta seção pelo uso de perfis falsos e/ou anônimos. Calcula-se que, no Twitter, aproximadamente 50 milhões de perfis (15% do total) são robôs projetados para simular pessoas reais (Varol et al., 2017), enquanto no Facebook, são mais 60 milhões segundo a própria empresa.[11] Uma das mais graves consequências dessa nova realidade é que, no Brasil e em diversos outros países, mídias sociais têm sido utilizadas com o intuito de manipular eleições nacionais.[12]

Nesse contexto, uma tarefa fundamental da CFD é a atribuição de autoria de mensagens curtas. Ela consiste em responder perguntas sobre a atribuição de um texto, usando apenas o conteúdo textual, ignorando qualquer outra informação como a autoria indicada pela mídia social, ou o grafo de relacionamentos com outros usuários. Vários são os possíveis cenários para essa tarefa, indo desde a atribuição de uma mensagem a um conjunto finito de suspeitos conhecidos (cenário *closed-set*), até a verificação de autoria em um cenário aberto (*open-set*), onde se precisa decidir se uma mensagem foi escrita por um suspeito ou por um desconhecido (Figura 5). No contexto de mídias sociais, a atribuição de autoria de mensagens curtas tem o potencial de identificar publicações e citações suspeitas.

No contexto apresentado na Figura 1, a etapa fundamental e mais desafiadora para o problema de atribuição de autoria de mensagens curtas é a de Mineração de Conteúdo. Nela está o desenvolvimento das técnicas de IA para a solução do problema baseados nos dados coletados nas mídias sociais.

O problema de atribuição de autoria têm sido estudado há bastante tempo no contexto de textos longos, como romances e artigos (Juola, 2008; Stamatatos, 2009; Koppel; Schler; Argamon, 2009). Contudo, o contexto de mídias sociais apresenta características novas e desafiadoras, como: mensagens muito curtas; maior variedade de símbolos (e.g., *emojis* e *emoticons*); gírias e erros de grafia (intencionais ou não); além de termos específicos das plataformas como menções a usuários e *hashtags*. Esse cenário tem sido recentemente endereçado, mas ainda sem resultados definitivos. Em Rocha et al. (2017), os autores apresentam uma revisão do problema e do uso de técnicas de IA nesse contexto. Atualmente, o estado da arte para o problema consiste no uso de modernos modelos de IA discriminativos como CNN (Theóphilo; Pereira; Rocha, 2019; Ruder; Ghaffari; Breslin, 2016; Shrestha et al., 2017), mas ainda com amplo espaço para melhora (Theóphilo; Pereira; Rocha, 2019).



Fonte: Elaboração própria.

Figura 6 – As técnicas de detecção automática de notícias falsas se baseiam em: verificação do *conteúdo* textual e visual da matéria; análise do *contexto* da notícia, como comentários e perfis de usuários que a compartilham; e análise da *rede de difusão* da notícia suspeita, investigando como ela se espalha pelas mídias sociais.

Desinformação e fake news

A detecção de *fake news* pode ser vista e modelada como uma tarefa de forense em que uma análise *post-factum* pode auxiliar para entender parte das visões falsas ou manipuladas dos fatos sobre um evento de interesse.

Um exemplo recente são os rumores difundidos sobre a pandemia do novo coronavírus (Covid-19). Informações imprecisas e falsas tais como recomendações médicas e teorias de conspiração estão se difundindo muito rapidamente. Como dito pelo presidente da Organização Mundial da Saúde (Ball; Maxmen, 2020): "Não estamos apenas combatendo uma epidemia, mas combatendo uma infodemia. Notícias falsas se espalham mais rápido e mais facilmente do que o vírus, e são igualmente perigosas". O problema de detecção de notícias falsas tem sido um dos tópicos mais discutidos recentemente pela comunidade científica. Para aproveitar essas informações, vários estudos propuseram e utilizaram *posts* de mídias sociais como fonte de dados para a detecção de notícias falsas. Os estudos podem ser divididos em pesquisas baseadas no conteúdo, contexto e na interconexão ou rede de difusão (Figura 6).

Nas técnicas *baseadas em conteúdo*, alguns trabalhos estão se concentrando no aproveitamento de informações textuais e visuais do conteúdo de notícias (Khattar et al., 2019; Zhou; Wu; Zafarani, 2020). Por exemplo, técnicas foren-

ses são úteis para detectar se uma determinada imagem ou vídeo foi manipulado ou se foi gerado artificialmente.

Outra abordagem utilizada é a adição de mais informações *baseadas no contexto*, tais como comentários e perfis de usuários. Em Jin et al. (2017), os autores propuseram uma abordagem de fusão multimodal que utiliza não apenas o conteúdo de texto e imagem de um *tweet*, mas o contexto social, incluindo *hashtags*, menções e *retweets*. Em um trabalho recente, Cui, Wang e Lee (2019) apresentaram uma abordagem que leva em conta os sentimentos dos usuários como possível contexto. A hipótese dos autores é que os sentimentos dos usuários em relação ao conteúdo da notícia pode indicar o grau de veracidade das mesmas. Em Schwarz, Theóphilo e Rocha (2020), os autores propõem um método de detecção de notícias falsas que utiliza três fontes de informação: postagens, comentários e notícias relacionadas. Em Wu e Rao (2020), os autores apresentaram uma solução adaptável entre postagens de notícias e comentários para detectar notícias falsas. Ao aproveitar informações de contexto social, os métodos de detecção de notícias falsas podem ser mais robustos do que os métodos baseados em conteúdo.

Além disso, estudar como as notícias falsas se espalham nas redes sociais também atraiu a atenção da comunidade científica. Foram propostas pesquisas analisando a *interconexão e redes de difusão* para entender os mecanismos virais de conteúdos suspeitos. No trabalho de Jang et al. (2018), foram estudados os padrões na divulgação de notícias verdadeiras e falsas. Em Shu, Wang e Liu (2019), os autores definiram o processo de divulgação de notícias como a tríplice relação entre o conteúdo da notícia enganosa, o veículo de difusão e seus usuários.

Finalmente, ainda persistem muitos desafios no processo forense (Figura 1) para a detecção de noticias falsas. A etapa de Coleta de Dados é um passo importante para que as técnicas baseadas em interconexão e difusão possam detectar padrões, como as comunidades e os propagadores mais influentes na rede. A etapa de Sanitização e Filtragem desempenha um papel importante para que as técnicas baseadas no conteúdo e no contexto possam caracterizar os tópicos de desinformação utilizando as informações relevantes para a análise. Por último, o grande desafio se encontra na etapa de Mineração de Conteúdo, que hospeda a difícil tarefa de determinar se uma notícia é falsa e a correspondente checagem.

Desinformação em artigos científicos

Como forma de impulsionar o conhecimento científico e se aproximar do público leigo, as redes sociais estão sendo cada vez mais usadas para a divulgação de artigos científicos. Como efeito colateral desse uso, esse tipo de publicação está sendo livremente interpretado pelo público, que por vezes acaba distorcendo as evidências, chegando a conclusões irreais pela ausência de rigor em suas interpretações (Scheirer, 2020; Scheufele; Krause, 2019). Em um caso noticiado pela BBC,[13] políticos norte-americanos e chineses disseminaram informações

não comprovadas sobre a origem do vírus causador da pandemia Covid-19. Em suas mensagens, eles se basearam em publicações não revisadas por pares que futuramente foram removidas pelos próprios autores.

Em 2005, com o objetivo de testar o sistema de revisão de publicações de algumas conferências, um grupo de cientistas do Massachusetts Institute of Technology (MIT) desenvolveu um software chamado *Sci-GEN*[14] capaz de gerar artigos que visam imitar esteticamente uma publicação científica. Desse modo, os artigos produzidos por essa ferramenta foram aceitos em conferências como *World Multiconference on Systemics, Cybernetics and Informatics (WMSCI)* e em revistas do editorial da *IEEE* e *Springer*,[15] mostrando, desde aquele período, uma problemática tanto para o processo de revisão de artigos quanto para a possibilidade da geração automática de publicações científicas.

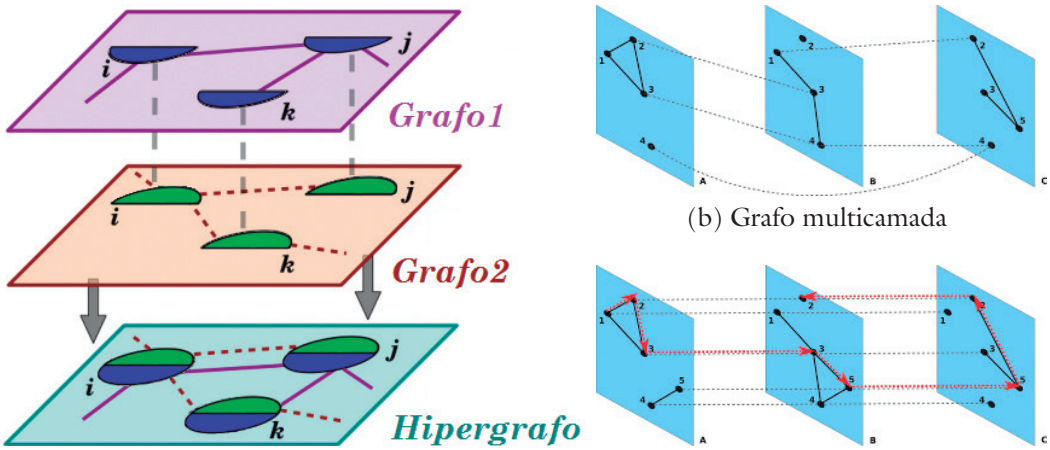
Tal problemática vem se agravando com o crescimento das fábricas de artigos (*paper mills*) e seu aperfeiçoamento de produção em massa. Qi, Zhang e Luo (2020) apresentam um avanço das técnicas de geração artificial de imagens científicas mediante sistemas de IA. Nessa pesquisa, imagens da área de biomedicina foram artificialmente geradas e mostradas para pesquisadores e especialistas da área, que não foram capazes de distinguir se as imagens eram artificiais ou reais. Marra et al. (2019) mostram que esses modelos de IA deixam artefatos no conteúdo de cada imagem gerada, que podem ser usados para identificar a origem sintética dessa imagem, possibilitando uma solução forense para esse problema.

O cenário atual da análise da integridade de artigos científicos ainda consiste na análise de suspeitas de fraudes isoladas, normalmente reportadas por leitores ou editores de conteúdo científico para organizações especializadas em integridade científica (e.g., *The Office of Research Integrity*[16]). Porém, com o crescimento de repositórios científicos abertos (e.g., ArXiv[17]), onde não há uma revisão sistemática das publicações, a análise de integridade científica está se direcionando para técnicas de IA de larga escala, visando minimizar que artigos fraudulentos desses repositórios sejam fontes de desinformação. Dessa forma, dentro do protocolo apresentado na Figura 1, os esforços mais recentes da área se concentram nas etapas de Filtragem e Mineração de Conteúdo. Na etapa de Filtragem, elementos suspeitos das publicações (i.e. textos e imagens) são selecionados por ferramentas capazes de processar milhares de artigos (Acuna; Brookes; Kording, 2018) que, durante a etapa de Mineração, são analisados por algoritmos treinados para detectar inconsistências ou plágio em seus conteúdos (Xiang; Acuna, 2020; Bucci, 2018).

Assim, os desafios futuros para se preservar a integridade científica, mitigar a expansão de mais fábricas de artigos e inibir que publicações falsas ecoem nas redes sociais dependerão de soluções forenses capazes de lidar com uma enorme quantidade de artigos publicados e que consigam identificar manipulações em imagens e textos científicos cada vez mais sofisticadas.

Quando o físico encontra o virtual

Os conteúdos digitais compartilhados, como imagens e *posts*, estão conectados entre si de diversas formas. Tais relações podem ser semânticas (por exemplo, o conjunto de *tweets* comentando a mesma notícia), mas também temporais e espaciais (por exemplo, todas as fotos capturadas na área externa do prédio do Capitólio no período da tarde). Na Figura 1, a etapa de organização semântica busca recuperar, a partir do conteúdo, algumas dessas informações e relações que podem acabar se perdendo à medida que os dados são compartilhados. Uma vez recuperadas, elas nos permitem explorar relações entre os dados que são fundamentais para auxiliar na compreensão do evento. Ao avaliar, por exemplo, como evoluíram a frequência e o número de publicações durante o evento da invasão do Capitólio, é possível reconstruir uma linha do tempo de como as pessoas se movimentaram pelas áreas próximas do prédio.[18]



(a) Hipergrafo (b) Grafo multicamada (c) Grafo multiplex

Fonte: Elaboração própria.

Figura 7 – (a) Uma rede de informação heterogênea que combina informações de duas camadas; (b) Um grafo multicamadas com links heterogêneos intra/inter camadas; (c) Um grafo multiplex com nós conectados com sua contraparte nas camadas adjacentes. A linha vermelha mostra um caminho possível para uma caminhada aleatória no grafo.

Apesar da importância de se explorar tais relações, essa não é uma tarefa trivial. Dados do mundo real, com atributos espaciais e temporais, têm propriedades diferentes em comparação com fontes não-relacionais de captura virtual (Atluri; Karpatne; Kumar, 2018), como imagens, vídeos, *posts*, *memes*, ou *hashtags* compartilhados sobre um evento. As instâncias não são independentes e distribuídas de forma idêntica, o que significa que, além da ligação com regiões espaciais e momentos no tempo, suas relações podem ser não estacionárias, ou seja, cada dado pode pertencer a diferentes populações ou classes que variam no tempo e no espaço.

Nesse contexto em que os dados provêm de fontes heterogêneas com dependência semântica, temporal ou espacial, a aplicação de abordagens tradicionais de mineração pode levar a um desempenho baixo e a interpretações incorretas (Jiang et al., 2015). Representações baseadas em grafos podem ser ferramentas úteis para a análise dessas relações (Boers et al., 2019; Ferreira et al., 2020). Entretanto, ainda há poucos estudos na CFD analisando dados multimodais, suas relações e a combinação de informações físicas e virtuais.

Dentre as abordagens existentes, os grafos multicamadas são ferramentas abrangentes, capazes de incorporar explicitamente múltiplos tipos de relações em uma modelagem adequada para se analisar fenômenos do mundo real e virtual (Boccaletti et al., 2014; Rossetti; Cazabet, 2018). Cada camada pode representar um conceito do mundo real ou virtual – como relações semânticas, momentos no tempo ou ligações espaciais – cujos nós podem representar instâncias de dados ou mesmo atores de interesse. Cada nó possui diferentes tipos de conexões e diferentes vizinhos em cada camada (Figura 7a), modelando relações a cada camada. Em uma rede social, podemos considerar, ao longo do tempo, as mudanças de relações entre *posts*, usuários e atores do evento à medida que são compartilhados (Figura 7c). Outras camadas podem representar a proximidade espacial dos dados compartilhados com pontos de interesse e de infraestrutura de uma cidade mostrando como o evento evoluiu espacialmente no mundo físico (Figura 7b).

Ainda que pouco exploradas pela comunidade de CFD, as modelagens em grafo tem potencial para complementar a análise individual do conteúdo dos dados gerados sobre um evento.

Desafios

Os métodos de IA apresentados atuam em etapas que precisam lidar com um grande volume de dados e automatizam processos intermediários que não necessitam da atenção exclusiva do perito. Entretanto, a inclusão dessas técnicas na análise forense também introduz desafios que acompanham a comunidade científica de IA nos últimos anos.

Interpretabilidade dos métodos

Os avanços da última década na área de aprendizado profundo (*Deep Learning*) impulsionaram o desenvolvimento de modelos estatísticos poderosos, geralmente baseados em redes neurais artificiais. Porém, à medida que a complexidade desses modelos aumentou – pela disponibilidade de dados e melhorias em hardware e nas técnicas utilizadas –, também se tornou mais difícil interpretar quais fatores influenciam suas decisões.

Essa limitação na interpretabilidade dos modelos reduz consideravelmente a sua aplicação em casos reais. Em contextos sensíveis – como cenários jurídicos, financeiros ou médicos – é essencial que as decisões de um algoritmo sejam transparentes e fundamentadas em uma lógica similar à humana. Nessa linha, a Lei Geral de Proteção de Dados (LGPD) no Brasil, inspirada pela Regulação Ge-

ral de Proteção de Dados (GDPR) da União Europeia, confere a um usuário o direito obter informações claras a respeito dos critérios e dos procedimentos utilizados no processo decisório de um algoritmo.[19] Sendo assim, para que os modelos sejam aplicáveis na prática, é fundamental que o seu fluxo de decisão seja interpretável não apenas para especialistas de IA, mas também usuários leigos.

A partir dessa necessidade, surgiu a área de Explainable Artificial Inteligence (XAI), levando pesquisadores a desenvolver formas de interpretar o que modelos complexos aprendem e como eles tomam suas decisões (Adadi; Berrada, 2018; Gilpin et al., 2018). Em Xie et al. (2020), os autores apresentam três vertentes principais de trabalhos voltados à interpretabilidade de redes neurais profundas. Os *métodos baseados em visualização* buscam apontar quais características dos dados de entrada – por exemplo, regiões de uma imagem ou palavras de um texto – tem influência na decisão do modelo (Selvaraju et al., 2017; Sundararajan; Taly; Yan, 2017). No contexto de análise de eventos, Padilha, Andaló e Rocha (2020) exploraram o uso dessas técnicas de visualização na organização cronológica de imagens do incêndio da Catedral de Notre-Dame. Os autores apontam que o modelo se baseia na evolução do incêndio e em detalhes estruturais da catedral para estimar quando cada foto foi tirada. Uma segunda linha de pesquisa alcança interpretabilidade a partir da *destilação de modelos*. Essa técnica envolve treinar modelos auxiliares que copiem as decisões de um modelo complexo feitas sobre um dado de entrada. Ao utilizar um modelo inerentemente interpretável como auxiliar – por exemplo, modelos lineares (Ribeiro; Singh; Guestrin, 2016), árvores de decisão (Frosst; Hinton, 2017), ou autômatos finitos (Hou; Zhou, 2020), procura-se imitar as decisões e qualidades do algoritmo mais complexo, porém de forma transparente e explicável. Finalmente, diversos trabalhos propõem *modelos intrinsecamente interpretáveis* que produzem uma explicação juntamente com a decisão tomada. Para isso, Zellers et al. (2019) e Hendricks et al. (2016) incluem uma tarefa adicional de produzir uma explicação textual que é otimizada simultaneamente à tarefa alvo do modelo.

Assegurar a interpretabilidade de modelo complexos de IA é um passo fundamental para que as decisões automatizadas ocorram de forma transparente e justa, permitindo que a sociedade aproveite os benefícios que esses modelos podem proporcionar.

Viés em dados e modelos

Os modelos complexos de IA são capazes de aprender diretamente de um grande volume de dados quais são as características mais eficientes para o problema em questão. Consequentemente, a sua qualidade está diretamente ligada à qualidade dos dados disponíveis. Como esses últimos são um recorte da nossa sociedade, muitas vezes eles acabam reproduzindo vieses indesejáveis de gênero, étnicos, econômicos, entre outros. Em razão de sua capacidade de aprendizado, os modelos acabam por capturar esses vieses, levando-os em consideração durante sua execução.

São vários os exemplos de sistemas reais que apresentam vieses com poder de impactar injustamente a vida das pessoas, como negar ofertas de emprego com base no sexo do candidato, aumentar o custo de um seguro de saúde de forma injusta, ou até mesmo influenciar juízes em decisões penais a partir da etnia do réu (Nelson, 2019). Em um episódio recente, foi constatado que um software utilizado no sistema penal americano para avaliar o risco de pessoas condenadas voltarem a cometer crimes apresentava claros indícios de viés contra pessoas negras.[20] No trabalho de Datta, Tschantz e Datta (2015), os autores apresentam um exemplo real em que um sistema de anúncios do Google sugeria menos vagas de empregos com altos salários para mulheres em comparação com homens.

Poderíamos afirmar que modelos enviesados podem existir desde que a decisão final seja sempre de um especialista humano. Contudo, a própria existência do resultado de um modelo que confirme vieses indesejados pode influenciar a decisão desse especialista e contribuir para a perpetuação do viés, reforçando o que Kahneman (2011) chama de "heurística de disponibilidade": exemplos que vêm de maneira natural e fácil em nossas mentes, têm o poder de inconscientemente influenciar decisões que são tomadas de forma aparentemente diligente.

Esses vieses podem se manifestar nos dados de diversas formas, como em atributos pessoais como a raça ou sexo de um indivíduo. A simples remoção desses atributos não garante a extinção do viés visto que eles podem estar fortemente correlacionados com outros atributos. Por exemplo, nos Estados Unidos e em diversos outros países, bairros têm alta correlação com a etnia de seus moradores e o uso desse atributo pode gerar modelos que tomam decisões considerando indiretamente a raça de uma pessoa.[21] Outra forma de manifestação diz respeito à representatividade dos diferentes grupos. O processo de coleta e anotação dos dados utilizados para treinamento dos modelos de IA pode gerar diferentes tipos de vieses: vieses de seleção, informação e detecção (Ntoutsi et al., 2020). O resultado pode ser a falta ou o excesso de representatividade de determinados grupos de pessoas que coincide com grupos sociais historicamente discriminados e que se reflete nas decisões desses modelos (Calders; Žliobaite, 2013).

A busca pela mitigação de vieses juntamente ao desenvolvimento de modelos mais interpretáveis, conforme detalhado antes, são desafios atuais e essenciais para a comunidade de IA e, de forma mais ampla, para a sociedade. Esses desafios não estão completamente desassociados, visto que modelos interpretáveis podem auxiliar no processo de busca por correções em modelos enviesados.

Vieses existem em nossas sociedades há muito tempo. Eles não foram criados pela tecnologia e nem serão resolvidos unicamente por ela. A solução para esse problema é bem mais ampla e exige um esforço multidisciplinar de todos nós, sem acreditar na panaceia de que, sozinha, a tecnologia tem a resposta para esta questão. O papel da tecnologia nesse caminho é auxiliar na identificação desses vieses e em sua mitigação.

Conclusões

O surgimento das mídias sociais na última década alterou a forma como nossa sociedade interage entre si, aumentando o alcance e a velocidade de propagação das informações. Todo acontecimento de grande visibilidade – como atentados terroristas e eventos de entretenimento – é registrado de inúmeras formas e instantaneamente compartilhado com o resto do mundo.

Nesse novo contexto, as informações provenientes de mídias sociais fornecem à CFD fontes de dados que, no geral, ultrapassam em volume as fontes tradicionais (e.g., veículos de rádio e televisão, testemunhas oculares). Em meio ao conteúdo propagado, há uma profusão de informações manipuladas e falsas, que influenciam a opinião das pessoas, além de ofuscar o real entendimento do que aconteceu. Aliados à imensa disponibilidade de dados, os avanços recentes em IA impulsionam o desenvolvimento de métodos capazes de gerar conteúdo realista automaticamente. Vivemos na era da pós-verdade; ao mesmo tempo em que não se pode confiar em tudo que é compartilhado, estamos envoltos em um oceano de informações que dificulta o rastreio de dados e fontes confiáveis sobre um acontecimento. Em meio a notícias falsas e desinformação, a análise forense se tornou mais complexa e desafiadora.

Neste artigo, discutimos como a popularização das mídias sociais e as inovações recentes em IA impactaram o processo de análise da CFD. Discutimos um protocolo forense adaptado a esse novo cenário que consiste na coleta de dados, sanitização e filtragem de relevância, organização semântica, e, por fim, mineração do conteúdo disponível. Essas etapas têm por objetivo recuperar e organizar os dados relevantes, de modo a permitir que o conhecimento e a experiência do perito sejam melhor aproveitados durante a análise.

Em razão do volume de informações que inviabiliza análises manuais em um tempo hábil, os métodos modernos de IA se tornaram ferramentas essenciais no arsenal do perito. A comunidade científica tem explorado com sucesso técnicas de aprendizado de máquina, visão computacional, e PLN em praticamente todas as etapas do processo de análise. Pela capacidade dos métodos em reconhecer e traçar relações entre os dados, sua aplicação complementa os esforços de um especialista, independente se o evento ocorreu no mundo físico, virtual, ou em ambos.

Apesar de serem ferramentas valiosas no processo forense, o uso dessas técnicas também traz consigo desafios que precisam ser discutidos para que sua aplicação seja viável em casos reais, especialmente em contextos sensíveis da CFD. Apesar de poderosos, modelos complexos existentes ainda são em grande parte "caixas-pretas" que não podem ser facilmente explicados. A interpretabilidade e transparência sobre o fluxo decisório desses modelos é uma característica essencial para seu uso na prática. Além disso, é necessário mitigar os vieses que muitas vezes são embutidos nos modelos durante o treinamento, para que suas decisões não os reproduzam.

Agradecimentos – Os autores agradecem os apoios financeiros da Fapesp (Proc. 2017/12646-3, DéjàVu), do CNPq (Proc. 304497/2018-5) e do Google Lara.

Notas

1 Disponível em: <https://www.darpa.mil/program/semantic-forensics>.

2 Disponível em: <https://www.darpa.mil/program/media-forensics>.

3 Disponível em: <https://www.unicamp.br/unicamp/sites/default/files/jornal/paginas/ju_578_paginacor_06e07_web.pdf>.

4 Disponível em: <https://revealproject.eu/>.

5 Disponível em: <https://www.nytimes.com/2021/01/09/technology/facial-recognition-clearview-capitol.html>.

6 Disponível em: <https://www1.folha.uol.com.br/cotidiano/2018/09/veja-o-passo-a-passo-da-noticia-falsa-que-acabou-em-tragedia-em-guaruja.shtml>.

7 Disponível em: <https://abcnews.go.com/blogs/technology/2013/04/boston-marathon-bombing-the-waves-of-social-media-reaction>.

8 Disponível em: <https://forensic-architecture.org/investigation/the-grenfell-tower-fire>

9 Disponível em: <https://www.bbc.com/news/world-us-canada-55578092>.

10 Disponível em: <https://www.nytimes.com/2015/06/07/magazine/the-agency.html>.

11 Disponível em: <http://www.businessinsider.com/facebook-raises-duplicate-fake-account-estimates-q3-earnings-2017-11>.

12 Disponível em: <https://www.bbc.com/portuguese/brasil-45666742>. Disponível em:<https://money.cnn.com/interactive/media/the-macedonia-story/>.

13 Disponível em: <https://www.bbc.com/news/world-52224331>.

14 Disponível em: <https://pdos.csail.mit.edu/archive/scigen>.

15 Disponível em: <https://news.mit.edu/2015/how-three-mit-students-fooled-scientific-journals-0414>.

16 Disponível em: <https://ori.hhs.gov/>.

17 Disponível em: <https://arxiv.org>.

18 Disponível em: <https://www.nytimes.com/interactive/2021/01/12/us/capitol-mob-timeline.html>.

19 Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm>.

20 Disponível em: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

21 Disponível em: <https://www.bloomberg.com/graphics/2016-amazon-same-day/>.

Referências

ACUNA, D. E.; BROOKES, P. S.; KORDING, K. P. *Bioscience-scale automated detection of figure element reuse*. Cold Spring Harbor Laboratory, fev. 2018.

ADADI, A.; BERRADA, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, v.6, p.52138-60, 2018.

ATLURI, G.; KARPATNE, A.; KUMAR, V. Spatio-temporal data mining: A survey of problems and methods. *ACM Comput. Surv.*, v.51, n.4, p.1-83, ago. 2018.

BALL, P.; MAXMEN, A. The epic battle against coronavirus misinformation and conspiracy theories. *Nature*, v.581, n.7809, p.371-4, 2020.

BOCCALETTI, S. et al. The structure and dynamics of multilayer networks. *Physics Reports*, v.544, n.1, p.1-122, 2014.

BOERS, N. et al. Complex networks reveal global pattern of extreme-rainfall teleconnections. *Nature*, v.566, n.7744, p.373, 2019.

BUCCI, E. M. Automatic detection of image manipulations in the biomedical literature. *Cell Death & Disease*, Springer Science and Business Media LLC, v.9, n.3, mar. 2018.

CALDERS, T.; ŽLIOBAITE, I. Why unbiased computational processes can lead to discriminative decision procedures. In: *Discrimination and privacy in the information society*. s.l.: Springer, 2013. p.43-57.

CASEY, E. *Digital evidence and computer crime: Forensic science, computers, and the internet*. s.l.: Academic Press, 2011.

CHEN, E.; LERMAN, K.; FERRARA, E. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, v.6, n.2, p.e19273, 2020.

CHESNEY, B.; CITRON, D. Deep fakes: a looming challenge for privacy, democracy, and national security. *California Law Review*, v.107, n.6, p.1753-820, 2019.

CHISUM, W. J.; TURVEY, B. Evidence dynamics: Locard's exchange principle & crime reconstruction. *Journal of Behavioral Profiling*, v.1, n.1, p.1-15, 2000.

CINELLI, M. et al. The covid-19 social media infodemic. *arXiv preprint*, arXiv:2003.05004, 2020.

CUI, L.; WANG, S.; LEE, D. Same: sentiment-aware multi-modal embedding for detecting fake news. In: *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. s.l.: s.n., 2019. p.41-8.

DATTA, A.; TSCHANTZ, M. C.; DATTA, A. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, v.2015, n.1, p.92-112, 2015.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd.*, v.96, n.34, p.226-31, 1996.

FERREIRA, A. et al. Counteracting the contemporaneous proliferation of digital forgeries and fake news. *Anais da Academia Brasileira de Ciências*, v.91, n.1, p.e20180149, 2019.

FERREIRA, L. N. et al. Spatiotemporal data analysis with chronological networks. *Nature Communications*, v.11, n.1, p.1-11, 2020.

FROSST, N.; HINTON, G. Distilling a neural network into a soft decision tree. *arXiv preprint*, arXiv:1711.09784, 2017.

GILPIN, L. H. et al. Explaining explanations: An overview of interpretability of machine learning. In: *IEEE* INTERNATIONAL CONFERENCE ON DATA SCIENCE AND ADVANCED ANALYTICS (DSAA). s.l.: s.n., 2018. p.80-9.

GUNNING, D. et al. XAI—Explainable artificial intelligence. *Science Robotics*, v.4, n.37, 2019.

GUPTA, A.; LAMBA, H.; KUMARAGURU, P. $1.00 per rt #bostonmarathon #prayforboston: Analyzing fake content on twitter. In: *2013 APWG eCrime Researchers Summit*. s.l.: s.n., 2013. p.1-12.

HENDRICKS, L. A. et al. Generating visual explanations. In: EUROPEAN CONFERENCE ON COMPUTER VISION (ECCV). s.l.: s.n., 2016. p.3-19.

HERNANDEZ-SUAREZ, A. et al. A web scraping methodology for bypassing twitter API restrictions. *arXiv preprint*, arXiv:1803.09875, 2018.

HOU, B.-J.; ZHOU, Z.-H. Learning with interpretable structure from gated RNN. *IEEE Transactions on Neural Networks and Learning Systems*, v.31, n.7, p.2267-79, 2020.

JANG, S. M. et al. A computational approach for examining the roots and spreading patterns of fake news: Evolution tree analysis. *Computers in Human Behavior*, v.84, p.103-13, 2018.

JIANG, Z. et al. Focal-test-based spatial decision tree learning. *IEEE Trans. Knowl. Data Eng.*, v.27, n.6, p.1547-59, 2015.

JIN, Z. et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA. s.l.: s.n., 2017. p.795-816.

JUOLA, P. Authorship attribution. *Foundations and Trends® in Information Retrieval*, v.1, n.3, p.233-334, 2008.

KAHNEMAN, D. *Thinking, fast and slow*. s.l.: Macmillan, 2011.

KHATTAR, D. et al. Mvae: Multimodal variational autoencoder for fake news detection. In: THE WORLD WIDE WEB CONFERENCE (WWW). s.l.: s.n., 2019. p.2915-21.

KOPPEL, M.; SCHLER, J.; ARGAMON, S. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, v.60, n.1, p.9-26, 2009.

LAMERI, S. et al. Who is my parent? reconstructing video sequences from partially matching shots. In: IEEE INTERNATIONAL CONFERENCE ON IMAGE PROCESSING (ICIP). s.l.: s.n., 2014.

LEFÈVRE, T. Big data in forensic science and medicine. *Journal of Forensic and Legal Medicine*, v.57, p.1-6, 2018.

MARRA, F. et al. Do GANs leave artificial fingerprints? In: IEEE CONFERENCE ON MULTIMEDIA INFORMATION PROCESSING AND RETRIEVAL (MIPR). s.l.: s.n., 2019.

MCFARLAND, D. A.; LEWIS, K.; GOLDBERG, A. Sociology in the era of big data: The ascent of forensic social science. *The American Sociologist*, v.47, n.1, p.12-35, 2016.

MIDDLETON, S. E.; PAPADOPOULOS, S.; KOMPATSIARIS, Y. Social computing for verifying social media content in breaking news. *IEEE Internet Computing*, v.22, n.2, p.83-9, 2018.

NELSON, G. S. Bias in artificial intelligence. *North Carolina Medical Journal*, v.80, n.4, p.220-2, 2019.

NGUYEN, D. T. et al. Automatic image filtering on social networks using deep learning and perceptual hashing during crises. *arXiv preprint*, arXiv:1704.02602, 2017.

NTOUTSI, E. et al. Bias in data-driven artificial intelligence systems—an introductory survey. *WIREs Data Mining and Knowledge Discovery*, v.10, n.3, p.e1356, 2020.

OMEZI, N.; JAHANKHANI, H. Proposed forensic guidelines for the investigation of fake news. In: *Policing in the Era of AI and Smart Societies*. s.l.: s.n., 2020. p.231-65.

PADILHA, R. et al. Unraveling the notre dame cathedral fire in space and time: an x-coherence approach. In: *A ser publicado em Crime Science and Digital Forensics*: A Holistic View. s.l.: CRC Press, 2021.

PADILHA, R.; ANDALÓ, F. A.; ROCHA, A. Improving the chronological sorting of images through occlusion: A study on the notre-dame cathedral fire. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP). s.l.: s.n., 2020. p.2972-6.

PADILHA, R. et al. Forensic event analysis: From seemingly unrelated data to understanding. *IEEE Security and Privacy*, v.18, n.6, p.23-32, 2020.

PINHEIRO, G. et al. Detection and synchronization of video sequences for event reconstruction. In: IEEE INTERNATIONAL CONFERENCE ON IMAGE PROCESSING (ICIP). s.l.: s.n., 2019.

POLLITT, M. A history of digital forensics. In: International Conference on Digital Forensics (IFIP). s.l.: s.n., 2010. p.3-15.

PRATES, M. O.; AVELAR, P. H.; LAMB, L. C. Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications*, p. 1-19, 2019.

QI, C.; ZHANG, J.; LUO, P. Emerging concern of scientific fraud: Deep learning and image manipulation. Cold Spring Harbor Laboratory, nov. 2020.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "Why should I trust you?" Explaining the predictions of any classifier. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. s.l.: s.n., 2016. p.1135-144.

ROCHA, A. et al. Authorship attribution for social media forensics. *IEEE Transactions on Information Forensics and Security*, v.12, n.1, p.5-33, 2017.

RODRIGUES, C. M. et al. Image semantic representation for event understanding. In: *IEEE International Workshop on Information Forensics and Security (WIFS)*. s.l.: s.n., 2019. p.1-6.

ROSSETTI, G.; CAZABET, R. Community discovery in dynamic networks: A survey. *ACM Comput. Surv.*, v.51, n.2, fev. 2018. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3172867>.

RUDER, S.; GHAFFARI, P.; BRESLIN, J. G. Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. *arXiv preprint*, arXiv:1609.06686, 2016.

SAMMES, T.; JENKINSON, B. *Forensic computing*. s.l.: Springer, 2007.

SCHEIRER, W. A pandemic of bad science. *Bulletin of the Atomic Scientists*, Informa UK Limited, v.76, n.4, p.175-84, 2020.

SCHEUFELE, D. A.; KRAUSE, N. M. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, v.116, n.16, p.7662-9, jan. 2019.

SCHNEIDER, M.; CHANG, S. A robust content based digital signature for image authentication. In: IEEE INTERNATIONAL CONFERENCE ON IMAGE PROCESSING (ICIP). s.l.: s.n., 1996. v.3, p.227-30, v.3.

SCHWARZ, S.; THEÓPHILO, A.; ROCHA, A. Emet: Embeddings from multilingual-encoder transformer for fake news detection. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP). s.l.: s.n., 2020. p.2777–81.

SELVARAJU, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. s.l.: s.n., 2017. p.618-26.

SHRESTHA, P. et al. Convolutional neural networks for authorship attribution of short texts. In: *Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. s.l.: s.n., 2017. p.669-74.

SHU, K.; WANG, S.; LIU, H. Beyond news contents: The role of social context for fake news detection. In: *ACM International Conference on Web Search and Data Mining (WSDM)*. s.l.: s.n., 2019. p.312-20.

SONG, L. et al. Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recognition*, v.102, p.107-73, 2020.

STAMATATOS, E. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, v.60, n.3, p.538-56, 2009.

SUNDARARAJAN, M.; TALY, A.; YAN, Q. Axiomatic attribution for deep networks. In: *International Conference on Machine Learning (ICML)*. s.l.: s.n., 2017. p.3319-28.

THEÓPHILO, A.; PEREIRA, L. A.; ROCHA, A. A needle in a haystack? Harnessing onomatopoeia and user-specific stylometrics for authorship attribution of micro-messages. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. s.l.: s.n., 2019. p.2692-6.

VAROL, O. et al. Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint*, arXiv:1703.03107, 2017.

VENKATESAN, R. et al. Robust image hashing. In: *IEEE International Conference on Image Processing (ICIP)*. s.l.: s.n., 2000. p.664-6.

WU, L.; RAO, Y. Adaptive interaction fusion networks for fake news detection. *arXiv preprint*, arXiv:2004.10009, 2020.

XIANG, Z.; ACUNA, D. E. Scientific image tampering detection based on noise inconsistencies: A method and datasets. *arXiv preprint*, arXiv:2001.07799, 2020.

XIE, N. et al. Explainable deep learning: A field guide for the uninitiated. *arXiv preprint*, arXiv:2004.14545, 2020.

YANG, F. et al. Asymmetric co-teaching for unsupervised cross-domain person re-identification. In: *AAAI*. s.l.: s.n., 2020. p.12597-604.

ZELLERS, R. et al. From recognition to cognition: Visual commonsense reasoning. In: *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. s.l.: s.n., 2019. p.6720-31.

ZHAI, Y. et al. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In: *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. s.l.: s.n., 2020. p.9021-30.

ZHOU, X.; WU, J.; ZAFARANI, R. Safe: Similarity-aware multi-modal fake news detection. *arXiv preprint*, arXiv:2003.04981, 2020.

RESUMO – A Ciência Forense Digital surgiu da necessidade de tratar problemas forenses na era digital. Seu mais recente desafio está relacionado ao surgimento das mídias sociais, intensificado pelos avanços da Inteligência Artificial. A produção massiva de dados nas mídias sociais tornou a análise forense mais complexa, especialmente pelo aperfeiçoamento de modelos computacionais capazes de gerar conteúdo artificial com alto realismo. Assim, tem-se a necessidade da aplicação de técnicas de Inteligência Artificial para tratar esse imenso volume de informação. Neste artigo, apresentamos desafios e oportunidades associados à aplicação dessas técnicas, além de fornecer exemplos de seu uso em situações reais. Discutimos os problemas que surgem em contextos sensíveis e como a comunidade científica tem abordado esses tópicos. Por fim, delineamos futuros caminhos de pesquisa a serem explorados.

PALAVRAS-CHAVE: Ciência forense digital, Inteligência Artificial, Aprendizado de máquina, Mídias sociais, Fake news.

ABSTRACT – Digital Forensics emerged from the need to perform forensic tasks in the digital age. Its most recent challenges are related to the popularization of social media and were intensified by the advance of Artificial Intelligence. The generation of massive social media data made forensic analyses more complex, mainly due to improvements in computational models able to artificially create highly realistic content. Because of this, Artificial Intelligence techniques have been studied and used to process the massive volume of information. This paper discusses the challenges and opportunities associated with such methods and provides real case examples, as well as the problems that arise when using these approaches in sensitive contexts and how the scientific community has approached these topics. Finally, it draws future research paths to be explored.

KEYWORDS: Digital forensics, Artificial Intelligence, Machine learning, Social media, Fake news.

Rafael Padilha é doutorando do Instituto da Computação da Universidade Estadual de Campinas (Unicamp). Contribuiu igualmente no desenvolvimento do artigo.
@ – rafael.padilha@ic.unicamp.br / https://orcid.org/0000-0003-1944-5475.

Antônio Theóphilo é doutorando do Instituto da Computação da Universidade Estadual de Campinas (Unicamp). Contribuiu igualmente no desenvolvimento do artigo.
@ – antonio.theophilo@ic.unicamp.br / https://orcid.org/0000-0003-1408-0745.

Fernanda A. Andaló é pesquisadora colaboradora do Instituto da Computação da Universidade Estadual de Campinas (Unicamp). @ – feandalo@ic.unicamp.br / https://orcid.org/0000-0002-5243-0921.

Didier A. Vega-Oliveros é pesquisador de pós-doutorado do Instituto da Computação da Universidade Estadual de Campinas (Unicamp). @ – davo@unicamp.br / https://orcid.org/0000-0001-9569-3775.

João P. Cardenuto é doutorando do Instituto da Computação da Universidade Estadual de Campinas (Unicamp). @ – phillipe.cardenuto@ic.unicamp.br / https://orcid.org/0000-0002-8370-6329.

Gabriel Bertocco é doutorando do Instituto da Computação da Universidade Estadual de Campinas (Unicamp). @ – gabriel.bertocco@ic.unicamp.br / https://orcid.org/0000-0002-7701-7420.

José Nascimento é doutorando do Instituto da Computação da Universidade Estadual de Campinas (Unicamp). @ – jose.nascimento@ic.unicamp.br / https://orcid.org/0000-0003-3450-6029.

Jing Yang é doutorando do Instituto da Computação da Universidade Estadual de Campinas (Unicamp). @ – jing.yang@ic.unicamp.br / https://orcid.org/0000-0002-0035-3960.

Anderson Rocha é professor associado do Instituto da Computação da Universidade Estadual de Campinas (Unicamp). @ – anderson.rocha@ic.unicamp.br / https://orcid.org/0000-0002-4236-8212.

Recebido em 22.2.2021 e aceito em 1º.3.2021.

[I-IX] Universidade Estadual de Campinas, Instituto da Computação, Campinas, São Paulo, Brasil.