CIÊNCIA DE DADOS E PRODUÇÃO DE CONHECIMENTOS DE INTELIGÊNCIA

POTENCIAL DA ANÁLISE DE DADOS DE REDES SOCIAIS DIGITAIS PARA A ATIVIDADE DE INTELIGÊNCIA

Daniel Fugisawa de Souza [*]
David Ricardo Damasceno do Bomfim [**]

Resumo

Nos últimos 20 anos, o escopo de interesse da atividade de Inteligência expandiu de modo a abranger conteúdos e métodos que foram gerados de forma digital (ou passaram a deixar registros digitais). Ante as peculiaridades dos dados de redes sociais, torna-se necessário incorporar técnicas da Ciência de Dados ao conjunto de métodos de análise de Inteligência. O presente artigo expõe o potencial das técnicas da Ciência de Dados na análise de dados de redes sociais digitais para a produção de conhecimentos de Inteligência. A motivação deste ensaio é divulgar e fomentar o debate acerca do aperfeiçoamento constante de métodos, técnicas e ferramentas da Ciência de Dados adotados na análise de redes sociais.

Palavras-chave: Inteligência, Ciência de Dados, produção de conhecimentos de Inteligência, análise de dados, redes sociais.

DATA SCIENCE AND INTELLIGENCE KNOWLEDGE PRODUCTION

POTENTIAL OF THE ANALYSIS OF SOCIAL NETWORKS FOR INTELLIGENCE ACTIVITY

Abstract

In the last 20 years, the scope of interests of the Intelligence analysis has broadened to encompass the contents and methods which were created digitally (or began to leave digital footprints). Considering the peculiarity of social networks data, it is necessary to incorporate Data Science techniques to the Intelligence analysis skill set. This article displays the potential of Data Science techniques in the analysis of social networks data for Intelligence knowledge production. The motivation behind this essay is to publicize and foster the debate concerning the constant improvement of Data Science methods, techniques and tools adopted in social networking analysis.

* Oficial de Inteligência

** Oficial de Inteligência

Keywords: *Intelligence, Data Science, Intelligence knowledge production, data analysis, social networks.*

CIENCIA DE DATOS Y PRODUCCIÓN DE CONOCIMIENTOS DE INTELIGENCIA

POTENCIAL DEL ANÁLISIS DE DATOS DE REDES SOCIALES PARA LA ACTIVIDAD DE INTELIGENCIA

Resumen

En los últimos 20 años, el ámbito de interés de la actividad de Inteligencia se ha ampliado para abarcar contenidos y métodos que se generaron digitalmente (o empezaron a dejar registros digitales). Dadas las peculiaridades de los datos de las redes sociales, es necesario incorporar métodos de la Ciencia de Datos al conjunto de métodos de análisis de Inteligencia. Este artículo expone el potencial de las técnicas de la Ciencia de Datos en el análisis de datos de redes sociales digitales para la producción de conocimiento de Inteligencia. La motivación de este ensayo es difundir y fomentar el debate sobre el perfeccionamiento constante de métodos, técnicas y herramientas de la Ciencia de Datos adoptados en el análisis de redes sociales.

Palabras clave: *Inteligencia, Ciencia de datos, producción de conocimiento de Inteligencia, análisis de datos, redes sociales.*

Introdução

A Ciência de Dados tem como objetivo revelar padrões e relações de fenômenos —não óbvios à percepção humana — a partir de dados para a produção de conhecimento para a tomada de decisão. (HAYASHI, 1996, e KELLEHER; TIERNEY, 2018). A Ciência de Dados resulta da combinação entre o desenvolvimento de análises com dados e a elaboração de novos algoritmos inteligentes capazes de unir a Estatística e a Ciência da Computação (BURLINGAME; NIELSEN, 2012).

A definição de Ciência de Dados a partir de seu objetivo guarda uma semelhança quase de identidade com a definição legal brasileira do ramo Inteligência da atividade de Inteligência. O Decreto nº 4.376/2002 define a função Inteligência como

> a atividade de **obtenção e análise de dados** e informações e de **produção e difusão de conhecimentos**, dentro e fora do território nacional, relativos a fatos e situações de imediata ou **potencial influência sobre o processo decisório**, a ação governamental, a salvaguarda e a segurança da sociedade e do Estado.

A definição legal do ramo Inteligência destaca a produção de conhecimentos relativos a situações de potencial influência sobre o Processo Decisório. A investigação de padrões e relações de fenômenos, bem como a elaboração de modelos preditivos, que pertencem ao conjunto de análises provenientes da aplicação de métodos da Ciência de Dados, permitem produzir conhecimentos de Inteligência. Essa semelhança reside em uma ideia comum a ambas as atividades: o conceito de ***actionable intelligence***.

No contexto da Ciência de Dados, *actionable intelligence* refere-se à capacidade de predição de cenários e de diagnóstico de uma situação que permita a ação de um decisor, com vistas a alcançar algum resultado desejado, e não uma mera descrição do passado e da atualidade. Na atividade de Inteligência, *actionable intelligence* apenas ressalta que o conhecimento a ser difundido para o cliente sirva de fato para a tomada de decisão. A principal expectativa do decisor para o produto de Inteligência, que destaca essa atividade das demais, guarda relação direta com a capacidade de apresentar perspectivas futuras, ou seja, predizer cenários que possam estimular ações no tocante ao Processo Decisório. Logo, a Ciência de Dados tem um grande potencial para oferecer um conjunto de métodos e técnicas essenciais para a produção de conhecimentos de Inteligência.

Este estudo apresenta possibilidades da aplicação de técnicas da Ciência de Dados na análise de dados de redes sociais, especialmente aquelas voltadas à capacidade de automação por mecanismos de Inteligência Artificial para a produção de conhecimentos de Inteligência.

A investigação sobre as possíveis aplicações das técnicas da Ciência de Dados para a satisfação das necessidades da Inteligência de Estado visa a determinar o impacto do uso das técnicas da Ciência de Dados na produção de conhecimentos da atividade de Inteligência, em especial na necessidade de análise de dados de redes sociais.

Este ensaio se justifica pela necessidade de consolidar, na comunidade brasileira de Inteligência, o aperfeiçoamento constante de métodos, técnicas e ferramentas de coleta e análise de dados, sobretudo aqueles coletados em plataformas digitais de rede social, que estão entre as fontes abertas mais volumosas em dados de interesse para a atividade de Inteligência. A otimização da análise das frações coletadas em fontes abertas confere eficiência na preparação e no processamento dos dados. Com a automação dessas etapas, o analista de Inteligência pode dedicar mais tempo de trabalho à aplicação de técnicas acessórias fundamentais de análise, ainda distantes de serem automatizadas integralmente por tecnologias de computação cognitiva[1].

Para cumprir com o objetivo exposto, conduziu-se uma pesquisa qualitativa de revisão sistemática da literatura — artigos científicos, livros e periódicos — sobre Ciência de Dados (em especial no contexto da *Big Data Analytics* [2]) e Teoria da Inteligência. A escolha da bibliografia obedeceu a três critérios: (1) relevância histórica (pela evolução conceitual ou pela atualidade e pela aplicabilidade do conhecimento do livro ou artigo); (2) relevância acadêmica medida pelo número de citações e (3) relevância para a doutrina de Inteligência e para a produção do conhecimento de Inteligência. Na pesquisa bibliográfica, investigaram-se aspectos conceituais e abordagens práticas da Inteligência de fontes abertas voltadas para a análise de dados de redes sociais digitais.

A hipótese de trabalho deste artigo é de que a associação de técnicas de Ciência de Dados é capaz de oferecer um novo paradigma para a produção de conhecimentos de Inteligência. Ao se considerar o volume e a variedade de dados a serem processados por meio de uma infraestrutura específica, seria possível estabelecer uma sequência de requisitos a serem satisfeitos em um *pipeline*[3] de dados que gere conhecimentos analíticos, prescritivos, explicativos ou preditivos, que transforma conhecimento em *actionable intelligence* para a tomada de decisão.

Na primeira seção, destaca-se a importância da adoção de métodos eficientes de reunião de dados de fontes abertas, com vistas a dedicar maior tempo do esforço do analista de Inteligência ao pensamento crítico com a aplicação de técnicas acessórias para a produção de conhecimentos. A maior fração dos dados coletados de plataformas digitais de redes sociais classifica-se no conceito

1 A computação cognitiva abrange áreas como Ciência da Computação, Ciência da Informação, cognição e inteligência no sentido de investigar as estruturas e os processos internos componentes do processamento de informações do cérebro e do funcionamento da inteligência natural (WANG *et alii*, 2010). A computação cognitiva se baseia na simulação das capacidades cognitivas humanas, a partir de algoritmos de Aprendizagem de Máquina herdados do campo da Inteligência Artificial, assim como na reprodução (imitação) do raciocínio humano (ANANTHANARAYANAN *et alii*, 2011).

2 *Big Data Analytics* abrange técnicas que tornam efetiva a mineração de uma quantidade expressiva de dados produzidos em alta latência e consumidos em diversos formatos, normalmente em tempo real. As técnicas buscam promover modelagem, visualização, predição e otimização.

3 *Pipeline* de dados é a sequência de fases do processamento de dados, iniciada pela ingestão de um fluxo de dados e seguida do processamento contínuo de várias etapas, que será abandonada quando a informação gerada esteja na forma desejada.

de *Big Data*[4]. A primeira seção termina com a constatação da necessidade de automação da coleta e da análise de dados para proporcionar melhores resultados à produção de conhecimentos.

A segunda seção inicia-se com uma exposição da atribuição legal da Inteligência nacional para a identificação de ameaças e oportunidades para o Estado brasileiro. Em seguida, desenvolve-se a ideia da coleta de dados como prioritária para alcançar essa finalidade e define o ramo da Inteligência de fontes abertas (*Open Source Intelligence* – OSINT) que trata de dados oriundos de redes sociais digitais, qual seja, *Social Media Intelligence* (SOCMINT). Essa seção termina com a discussão acerca da importância de acompanhar o mais amplo espectro de objetos de interesse, com vistas a manter a vantagem competitiva ante os demais atores que difundem informações para contribuir com o Processo Decisório Nacional.

A terceira seção expõe algumas das principais técnicas da Ciência de Dados para a análise de dados de redes sociais aplicáveis à produção de conhecimentos de Inteligência. Entre as modalidades analíticas de dados capazes de suprir as necessidades da Inteligência, lista-se (de forma não-exaustiva) a léxica, a de conexões em rede social (*Social network analysis*), a de sentimentos, a de georreferenciamento e a de geoinferência. Esta parte finda com a apresentação das técnicas de automação do processamento de dados organizados no repositório de dados, em representação a frações de interesse coletadas, com foco nos mecanismos de Inteligência Artificial.

A quarta seção aponta algumas abordagens práticas de inteligência de dados de redes sociais digitais viabilizadas por técnicas da Ciência de Dados. Os exemplos abarcam aplicações no acompanhamento da ação de objetos de interesse para a atividade de Inteligência e na satisfação de consciência situacional[5] acerca de eventos ou estados de coisas que sugiram ameaça ou oportunidade para a manutenção da ordem social e para a segurança do Estado e da sociedade. Aplicações que subsidiam o combate ao terrorismo, à espionagem e à interferência externa são apresentadas.

Por fim, os autores do estudo se posicionam acerca da hipótese de trabalho, a partir das evidências, conforme apreciação das abordagens práticas de inteligência de dados de redes sociais digitais que empregaram técnicas da Ciência de Dados para solucionar demandas comuns no cotidiano da atividade de Inteligência. Conclui-se com considerações sobre o impacto de outras novas tecnologias no

4 *Big Data* é uma nova geração de tecnologias e arquiteturas projetadas para extrair economicamente valor de volumes muito grandes de uma ampla variedade de dados, permitindo a captura, descoberta e análise em alta velocidade (GANTZ; REINSEL, 2011).

5 Consciência situacional envolve compreender como informações, eventos e ações podem impactar o estado de coisas de um ambiente crítico, tanto imediatamente quanto prospectivamente. A consciência situacional favorece a tomada de decisão com uma visão do evento ou da conjuntura na forma de um sistema de informações, com entradas e saídas, em que se pode perceber como o ajuste de variáveis do ambiente pode adequar esse "sistema" com antecipação para evitar resultados indesejáveis. É especialmente vantajoso valer-se do conceito em áreas complexas e dinâmicas, como no controle de tráfego aéreo, em operações de infraestruturas críticas, em centrais de comando e controle e em serviços de emergência.

âmbito da análise de dados de redes sociais e sobre o aperfeiçoamento do uso sistemático de técnicas da Ciência de Dados para subsidiar a atividade de Inteligência.

Contextualização

Os métodos e técnicas necessários para se alcançar os objetivos da produção de conhecimentos de Inteligência de Estado são complexos porque são multidisciplinares. O desenvolvimento da atividade de Inteligência no âmbito da Administração Pública gera alguns desafios, especialmente no tocante à sensibilização da alocação de recursos para a aquisição de novas tecnologias, ao treinamento de pessoal, à adequação das expectativas em relação a prazos e entregas, além de outras especificidades necessárias para a adesão a iniciativas que sugiram o rompimento de paradigmas organizacionais.

O diálogo transparente entre os gestores de nível estratégico, em acesso permanente à cúpula do Estado, e os de nível tático, responsáveis pela prospecção de novas tecnologias, é essencial para viabilizar a aquisição de produtos e para a contratação de serviços necessários a sustentar a infraestrutura de soluções. As ações de sensibilização direcionadas aos diretores do órgão de Inteligência devem revelar os benefícios para a estratégia institucional, de modo a apontar como novos modelos de soluções podem aperfeiçoar o cumprimento da missão institucional.

Essas iniciativas organizacionais ganham importância especialmente quando sugerem ajustes de métodos e técnicas de trabalho, o que é comum no surgimento de novas tecnologias. Além de técnicas acessórias fundamentais para a avaliação e para a prospecção de cenários, como a Análise Estruturada[6] e a apreciação de vieses cognitivos, cabe, portanto, a capacitação tecnológica ante a evolução incessante dos métodos, técnicas e ferramentas computacionais de análise de dados.

O trabalho diferenciado do profissional de Inteligência deve abranger principalmente o exercício do pensamento crítico, a fim de desenvolver conhecimentos com base tanto em juízos quanto em raciocínios aprofundados e aptos a interpretar ou a prospectar eventos. Coleta e processamento de frações de fontes abertas ou obtidas por meio de buscas[7] operacionais devem ser automatizados na medida do possível. A produção de conhecimentos corre o risco de se limitar à elaboração de descrições de eventos baseadas unicamente em juízos, especialmente devido à escassez de tempo para a satisfação da oportunidade

6 Análise estruturada não se confunde com análise de dados estruturados. Trata-se de gênero que comporta uma variedade de técnicas. Análise estruturada refere-se a métodos de organizar e estimular o pensamento sobre problemas de Inteligência, na intenção de mitigar o impacto de vieses cognitivos (AMBROS; LODETTI, 2019). Já análise de dados estruturados refere-se a métodos e técnicas analíticas em geral para extrair valor a partir de um conjunto de dados estruturados, ou seja, dados representados na forma de uma estrutura rígida bem definida, a exemplo de dados de agregados macroeconômicos, como dados de balança comercial, de taxas de câmbio e de Produto Interno Bruto.

7 Busca é a ação especializada para obtenção de dados negados, mediante o emprego de técnicas operacionais (BRASIL, 2016b, p. 86).

do conhecimento a ser difundido para o tomador de decisão.

O aperfeiçoamento do processamento de frações coletadas em fontes abertas exige o emprego de sistemas cognitivos com mecanismos de Inteligêngia Arficial, como Aprendizado de Máquina e Processamento de Linguagem Natural.

Inteligência Artificial pode ser entendida como um grande conjunto de ferramentas para fazer computadores se comportarem de forma "inteligente" e de forma automatizada. Isto inclui assistentes de voz, sistemas de recomendação, carros autônomos[8].

Aprendizado de Máquina (*Machine Learning* – ML) é o campo de estudo que "dá a computadores a habilidade de aprender sem serem explicitamente programados", de acordo com Andrew Ng, cofundador do Google Brain, ex-cientista-chefe do Baidu, e professor do curso de *Machine Learning* pela Universidade de Stanford. Em ML, computadores aprendem padrões a partir de dados existentes, tal qual em modelos estatísticos tradicionais. No entanto, em análises estatísticas, o analista busca o modelo estocástico (probabilístico), dentro de um conjunto de modelos conhecidos, que poderia ter gerado os dados de fato observados[9].

Processamento de Linguagem Natural (*Natural Language Processing* – NLP) refere-se ao ramo da Ciência da Computação — mais especificamente um ramo da Inteligência Artificial — que objetiva dotar computadores da habilidade de compreender texto escrito e áudio emitido o mais próximo possível de como as pessoas o fazem. NLP combina linguística computacional (modelagem de linguagem humana, baseada em regras) com Aprendizado de Máquina[10].

Assim, a coleta[11] e a mineração[12] de grande parte das frações de interesse em fontes abertas de dados seriam obtidas e analisadas em quantidade e em

8 Em uma definição academicamente mais rigorosa, Inteligência Artificial (IA) possui quatro abordagens que a definem e a caracterizam na atualidade. As abordagens advêm das possíveis combinações entre as dimensões Comportamento *versus* Pensamento e Humano *versus* Racional, que são consideradas para se avaliar o resultado apresentado por um sistema ou por um equipamento de IA. A questão da racionalidade também pode considerar processos de pensamento para tomada de decisão (caracterização interna) ou comportamentos inteligentes desempenhados (caracterização externa). Os métodos utilizados por cada uma das quatro abordagens de IA consideram que a busca pela simulação ou reprodução da inteligência humana pode se desenvolver sob uma perspectiva de ciência empírica relacionada à Psicologia (observações e hipóteses sobre o comportamento humano e processos de pensamento) ou sob um viés racionalista (combinação de Matemática e Engenharia, com aportes de Estatística e Economia). Entre as abordagens de IA, aquela dedicada ao desenvolvimento de um agente racional é a que tem prevalecido na maior fração dos estudos e aplicações, por ser a mais genérica (flexível) e a mais acessível ao desenvolvimento científico. Em suma, projetos de IA têm se concentrado principalmente na construção de agentes que tomem decisões, mesmo sob incerteza, e apresentem o melhor resultado esperado, conforme as percepções captadas do ambiente onde esse agente opera (RUSSELL; NORVIG, 2021).

9 O Aprendizado de Máquina envolve a utilização de algoritmos para extrair informações de dados brutos e representá-los por meio de um modelo matemático que se ajusta a novas circunstâncias e que detecta e extrapola padrões (RUSSELL; NORVIG, 2021).

10 Processamento de Linguagem Natural consiste no desenvolvimento de modelos computacionais para a interação com sistemas computadorizados, que dependam de informações expressas em linguagem humana como insumo para processamento (RUSSELL; NORVIG, 2021).

11 Coleta é a ação especializada para a obtenção de dados de livre acesso (BRASIL, 2016b, p. 87).

12 Mineração é o processo de extração de padrões e conhecimentos de interesse a partir da coleta de um amplo volume de dados oriundos de repositórios de informação ou mesmo de um fluxo contínuo de dados (*stream*) (HAN *et alii*, 2012).

velocidade compatíveis com a produção (pela sociedade) e com a necessidade (do tomador de decisão), de forma consistente, por analistas de Inteligência versados nessas técnicas. A associação dessas ferramentas favorece o analista de Inteligência com uma maior fração de tempo disponível para a aplicação de técnicas acessórias fundamentais de análise, ainda distantes de serem automatizadas integralmente por tecnologias cognitivas.

De fato, o uso pervasivo da internet, de dispositivos que geram dados sobre indivíduos e organizações em tempo real, e o uso constante de redes sociais fazem com que a comunidade de Inteligência considere o impacto desses novos meios de sociabilidade não apenas em seus métodos de análise, mas também no bojo de objetos de interesse. Como se verifica a seguir, a transformação é tão radical nas formas de interações sociais que temas, técnicas, finalidades e tipos de produtos precisam ser repensados para dar conta das mudanças.

Na atualidade, *petabytes* diários de dados de fontes abertas de interesse da atividade de Inteligência são gerados em plataformas *on-line* de rede social. Os dados gerados por meio das plataformas de rede social apresentam características compatíveis com o que se tem por *Big Data*: dados gerados em expressivos volume, velocidade e variedade.

Plataformas de redes sociais adotam modelos de negócios que visam tanto a promover serviço para anunciantes de campanhas publicitárias, a exemplo do emprego de *Adwords*[13], quanto a prestar serviços de comunicação e de publicação de conteúdo, com o objetivo primordial de aumentar as adesões à plataforma. Na medida em que se consolida essa fidelização, sistemas de recomendação e recursos de divulgação da rotina de personalidades em evidência operam de modo a consumir o máximo de tempo e dedicação consciente dos usuários. Nesse contexto, desenvolve-se o *microtargeting,* ou seja, propaganda definida para difusão a um alvo determinado, conforme as preferências, insatisfações e perspectivas por ele externadas em perfil de rede social.

Por conta dessa natureza de mercantilizar as informações dos indivíduos e a previsão do comportamento dessas pessoas, resulta que gerar e reunir cada vez mais informações sobre os indivíduos é próprio da natureza das redes sociais digitais. Para a atividade Inteligência, essa característica tem o potencial de tornar essa mídia como a principal fonte de dados e informações para a produção de conhecimentos.

Comentários e opiniões são tratados para se reconhecer padrões a partir de modelos probabilísticos com algoritmos de Aprendizado de Máquina. Na medida em que opiniões, ideologias e preconceitos são transformados em alguma fórmula matemática, ganham um grau de confiabilidade e naturalização capaz de suprimir o questionamento sobre o que fornecem, sobre como operam e sobre os efeitos que produzem. Com isso, as

13 *Adwords* é o mecanismo que posiciona a campanha publicitária em espaços reservados na interface e dão maior destaque nos resultados de busca.

próprias opiniões e formas de pensamento são reduzidas em modelos que, além de coletar informações, reproduzem percepções e ideologias contidas em dados do passado — usados na criação do modelo (O'NEIL, 2017).

A intensidade da interação de usuários em rede favorece a compreensão de fenômenos e de comportamentos. Isso pode ser feito de forma mais eficiente por meio da concepção de sistemas inteligentes, capazes de analisar as circunstâncias de eventos, de elaborar a predição de cenários e de acompanhar as ações de alvos.

Logo, as plataformas de redes sociais são uma das fontes de dados mais férteis para a produção de conhecimentos de interesse da atividade de Inteligência.

Os métodos de pesquisa acerca da automação da extração, da categorização, do agrupamento, da sumarização e da indexação de informações têm se adaptado ao tratamento dinâmico dos dados de redes sociais em sortidos formatos, a fim de explorar todas as potencialidades das plataformas digitais de interação social.

A análise de dados de redes sociais digitais, especificamente para a atividade de Inteligência, apresenta-se como um desafio, visto que as plataformas são incompatíveis com uma estrutura hierárquica que controle a repercussão das postagens dos usuários. A organização de Inteligência tende a se beneficiar dessa liberdade, uma vez que a automação da seleção de frações de interesse para a análise de Inteligência pode facilitar o trabalho de inteligência de fontes abertas.

Inteligência baseada em dados de redes sociais digitais: oportunidades e vantagem competitiva

Consoante o disposto no artigo 4º da Lei nº 9.883/1999, entre as competências da Agência Brasileira de Inteligência (Abin) estão: planejar e executar ações, inclusive sigilosas, relativas à obtenção e à análise de dados para a produção de conhecimentos destinados a assessorar o Presidente da República; e avaliar as ameaças, internas e externas, à ordem constitucional.

De acordo com o Decreto nº 8.793/2016, que fixa a Política Nacional de Inteligência (PNI), cumpre à Inteligência Nacional o acompanhamento e a avaliação da conjuntura interna e externa, para buscar identificar fatos ou situações que possam resultar em ameaças, riscos ou oportunidades aos interesses da sociedade e do Estado. Conforme previsão do inciso VII do artigo 15 do Decreto nº 10.445/2020, uma das competências do Centro de Inteligência Nacional (CIN) da Abin abrange planejar, coordenar e implementar a produção de inteligência corrente e a "coleta estruturada de dados".

Cabe à Inteligência alertar o chefe de Estado acerca de ameaças e de oportunidades para a segurança nacional, assim como informar sobre tendências e prospectar cenários futuros. A Inteligência fornece subsídios para o Processo Decisório Nacional, no tocante à segurança

da sociedade e do Estado, por meio da difusão de conhecimentos. A produção de conhecimentos de Inteligência abrange o fornecimento de informações acuradas com imediato processamento e com célere difusão, a fim de cumprir com a oportunidade do conhecimento produzido e, ao menos, manter a vantagem competitiva ante outras organizações ou nações.

Das fases da produção do conhecimento de inteligência, a etapa da coleta é a que fornece o dado, a partir do qual se elabora o conhecimento. Sem dado, sem conhecimento. A coleta é o alicerce de todos os trabalhos de Inteligência e, portanto, qualquer produção conduzida sem fundamento em informações não passaria de um mero exercício de ilação. Cabe verificar a ocorrência de vieses, pois um algoritmo treinado a partir de um conjunto enviesado de dados repetirá um resultado com semelhante viés (*garbage in, garbage out*[14]). A coleta na atividade de Inteligência define-se como a aquisição de informações por meio de vários métodos que satisfaçam as demandas do Estado afetas à segurança nacional (LOWENTHAL, 2019).

Entre os métodos de captura de insumos para processamento analítico pela Inteligência, está a busca de dados em fontes abertas, especialidade conhecida como Inteligência de fontes abertas (OSINT). Essa técnica, que abrange a coleta de dados que possam ser obtidos legalmente a partir de repositórios de dados públicos, refere-se preponderantemente na atualidade a informações disponíveis na Internet.

Postagens em fóruns de discussão virtuais, arquivos digitais nos formatos texto e multimídia, metadados[15] de diversas ordens, informações técnicas (endereços *Internet Protocol* - IP, endereços *Domain Name System* - DNS, registros Whois), informações de georreferenciamento e plataformas de redes sociais são alguns dos recursos disponíveis para coleta *on-line*. Publicações oficiais, bases de dados governamentais, relatórios financeiros sobre empresas publicados em sites e outras bases de acesso restrito pela Internet, com acesso franqueado após acordos de cooperação, também configuram como fontes de obtenção de frações de interesse. Há de se considerar ainda, para o acompanhamento de ameaças, as bases de dados ilegalmente distribuídas por meio de plataformas de compartilhamento ponto a ponto (*Peer-to-peer* - P2P) ou divulgadas em fóruns virtuais (como 4chan[16] e 8kun[17]) e em sites da *Deep web*[18].

14 Metáfora utilizada na área de Computação para apontar que, se os dados que você usa como entrada de um sistema são "lixo", a saída (o resultado do processamento) semelhantemente será "lixo".

15 Metadados são informações de valor agregado para organizar, descrever, rastrear e melhorar o acesso a objetos de informação e a itens físicos e coleções, relacionados a esses objetos (GILLILAND, 2016).

16 https://www.4chan.org

17 https://8kun.top

18 *Deep Web* abrange páginas de textos, arquivos e demais recursos dispostos na *World Wide Web* cujos motores de buscas em geral não conseguem adicionar aos índices de páginas Web, de forma que consigam ser relacionados no resultado da pesquisa textual por meio de interface do motor de busca. As páginas eletrônicas da *Deep Web* possuem criptografia própria e são acessíveis apenas por meio de *softwares* específicos que decifram o caminho onde realmente está hospedado o recurso Web, como o *The Onion Routing* (TOR) e a plataforma *Freenet* (SHERMAN; PRICE, 2001).

Não só órgãos de Inteligência, mas também organizações privadas têm dedicado recursos para desenvolver capacidades de coleta de dados abertos de amplo acesso, pois precisam traçar estratégias para competir ou colaborar com outras instituições. Há ferramentas que permitem a busca e a extração de dados nos mais sortidos formatos em repositórios, desde servidores de arquivos e de correio eletrônico até ambientes com acesso dificultado como é na *Deep web*.

Tornou-se comum classificar a inteligência de mídias sociais como uma nova modalidade derivada da OSINT, devido a sua importância quanto ao potencial de obtenção de dados e a ampla miríade de métodos e técnicas empregados para processar dados não-estruturados coletados a partir das interações em plataformas de redes sociais. Essa modalidade tecnológica complementar, para processar *Big Data* em favor da tomada de decisão a partir do monitoramento de redes sociais digitais, foi definida como *Social Media Intelligence* (SOCMINT).

SOCMINT abrange a observação e a análise de indivíduos e grupos para compreender aspectos comportamentais afetos às relações e aos sistemas estabelecidos, o que demanda a avaliação conjugada de atitudes, de conjunturas e de características culturais para uma análise eficiente em prol da melhor tomada de decisão, especialmente quando há envolvimento de medidas de prevenção e repressão a distúrbios da ordem pública. SOCMINT pode tanto trazer oportunidades quanto provocar reveses para a segurança nacional e para os interesses estratégicos do Estado.

Entre essas oportunidades estão: a) o acompanhamento das informações difundidas pela população na ocasião de um evento de interesse das agências governamentais, a exemplo de catástrofes naturais, surtos epidêmicos ou manifestações sociais, sob um modelo de colaboração coletiva (*crowdsourcing*); b) a revelação do modo de articulação de organizações criminosas e de vertente extremista, a exemplo de ações para radicalização ou difusão de ideias de violência em desfavor da ordem pública; c) a promoção da consciência situacional em tempo real, de modo que o governo venha a prestar a melhor experiência aos usuários dos serviços públicos; d) a percepção dos cidadãos acerca dos resultados de políticas públicas; f) a atuação de grupos de ódio e de difusão de *fake news*; e g) a avaliação da interferência externa de países estrangeiros.

Entre os atores que conduzem ameaças por meio de plataformas de redes sociais digitais estão: a) terroristas e grupos extremistas que se valem de ferramentas *on-line* para perpetrar campanhas de desinformação e promoção do medo coletivo (elemento de guerra psicológica), propaganda ideológica, recrutamento de membros, difusão de informações acerca de fabricação de artefatos explosivos e sobre locais de interesse para perpetração de atentados; b) organizações criminosas que articulam e patrocinam atividades ilícitas, a exemplo de pornografia infantil, contrabando, tráfico de pessoas, lavagem de dinheiro e transações em moeda digital

em favor do tráfico de entorpecentes; c) grupos que articulam movimentos paredistas para causar instabilidade no fornecimento de serviços essenciais; d) negociação de dados pessoais sensíveis (cadastrais, médicos e financeiros, por exemplo); e e) organizações estatais e agentes privados patrocinados por nações concorrentes ou adversárias, que conduzem campanhas *on-line* para a desestabilização de instituições de Estado.

Outro aspecto que sugere a importância do investimento massivo em tecnologias que tratem dados coletados de plataformas digitais de rede social é a existente competição entre órgãos nacionais de Inteligência para se posicionarem como autoridade nas diversas temáticas que envolvem a segurança nacional. A Inteligência de Estado não concorre somente com frações governamentais congêneres para apresentar um panorama ou uma perspectiva sobre um evento ou uma ameaça com ineditismo e com a esperada oportunidade. Atores como a imprensa, os *think tanks*, o setor privado e mesmo contatos de ordem pessoal podem figurar como concorrentes do órgão de Inteligência de Estado, como também suscitar questionamentos acerca da eficiência dos meios oficiais de produção de conhecimentos estratégicos.

Há momentos em que é cabível a aplicação de técnicas tradicionais de coleta (que submetem alguém a uma entrevista, por exemplo) para a confecção de relatórios detalhados com interpretações e prospecções de cenários; entretanto, há emergências em que mais vale o emprego de tecnologias que busquem relações ou padrões não-óbvios de fenômenos e que alimentem em tempo real um painel (*dashboard*) acessível para consulta *on-line* pelo tomador de decisão.

Nessas emergências, é comum que cause mais impacto para o tomador de decisão o ineditismo da notícia apresentada do que a entrega de um relatório detalhado sobre o evento. O emprego de sistemas cognitivos dotados de algoritmos de Aprendizado de Máquina vem satisfazer tanto a inferência de relações entre variáveis para visualização abrangente do problema, quanto a capacidade preditiva para traçar estimativas.

Com a emergência das agências globais de notícia, a manutenção da consciência situacional pela Inteligência acerca de algum evento tende a ficar defasada na ordem de horas ou dias em relação à conquistada pela imprensa. Esta, por sua vez, tende a ficar, no mínimo, horas desafada em relação às primeiras postagens por perfis de pessoas ou instituições em plataformas de redes sociais. Por consequência, os órgãos governamentais de Inteligência tendem a ficar cada vez mais defasados e desacreditados caso não apostem com robustez no emprego das técnicas e no desenvolvimento das capacidades de SOCMINT.

A Primavera Árabe (dezembro de 2010) é um caso emblemático. O evento foi substancialmente articulado por meio de plataformas de rede social. Nessa situação, ativistas revelaram que usaram o Facebook para agendar os protestos, coordenaram as ações durante os protestos via Twitter

e difundiram as transmissões ao vivo por meio do Youtube. Enuncia-se o evento como o primeiro a proporcionar a difusão de informações sobre a manifestação com célere e generalizada violação das restrições governamentais.

Outros eventos relevantes para demonstrar o impacto das redes sociais no contexto da Inteligência abrangem: a) os protestos pela prisão do presidente deposto das Filipinas em abril de 2001; b) os ataques terroristas em Mumbai em novembro de 2008; c) os protestos em contestação aos resultados das eleições na Moldávia em abril de 2009; d) a suspeita de interferência russa nas eleições estadunidenses em 2016 e nas francesas em 2017; e e) a articulação em redes sociais da paralisação dos caminhoneiros autônomos ("crise do diesel") no Brasil em maio de 2018.

Na maioria dessses eventos, apesar de as agências de Inteligência e de as forças de segurança terem reagido tardiamente, vários atores governamentais passaram a estudar a aquisição de ferramentas e a contratação de profissionais qualificados para adotar soluções de coleta e de processamento de dados de redes sociais digitais. Esse esforço é referenciado como *dataveillance*[19] e compreende o monitoramento de organizações criminosas, de grupos extremistas e de atores que se prestam a constranger a ordem pública (CLARKE, 1988).

Técnicas de Ciência de Dados para tratamento de dados de redes sociais

SOCMINT vale-se de diversas modalidades de coleta e análise, que são, por vezes, até combinadas, para processar dados advindos de plataformas digitais de rede social. Embora a maioria dessas técnicas seja classificada como de análise textual, visto que dados em formato textual se apresentam em maior volume e exigem menor carga de processamento, cabe apresentar também técnicas especializadas para a análise de dados geoespaciais, de conexões em rede e de imagens. Entre as modalidades de SOCMINT mais usadas na coleta, estão a técnica de recuperação de informações e o emprego de *web crawlers*[20] e de *web scrapers*[21].

A recuperação ou a extração de informações abrange tanto a pesquisa de uma simples palavra-chave (que venha a denotar uma atividade guiada pelos grupos de interesse da atividade de Inteligência), como também a detecção automática de tópicos, a fim de classificar um texto relacionado a um possível diálogo entre terroristas ou entre lideranças de uma organização

19 *Dataveillance* é uma forma de vigilância contínua através do uso de dados e de metadados (CLARKE, 1988).

20 *Web crawlers* são algoritmos que indexam páginas da web em cascata, ou seja, dado um endereço inicial (ou conjunto de endereços iniciais) e algumas condições (por exemplo, quantos *links* ainda faltam, tipos de arquivos a serem ignorados), eles fazem o mapeamento de tudo o que está vinculado a partir do ponto de partida. Podem ser utlilizados para arquivamento de dados que estariam inativos em um sistema, mas que seriam úteis, por exemplo, para fins de auditoria e de processos afetos à verificação de conformidade (*compliance*).

21 *Web scrapers* são algoritmos que extraem conjuntos de dados de recursos *online* e os armazenam em um formato estruturado (XML ou planilha, por exemplo) para viabilizar ou facilitar o processamento posterior por ferramentas de análise de dados. O termo "raspagem de dados" é amplamente difundido para referenciar a ação de *web scrapers*.

criminosa, por exemplo. A informação é obtida por meio de um *software* dotado de algoritmos baseados em regras. É possível utilizar uma ferramenta de análise textual e um conjunto de dicionários para capturar e apreciar narrativas em fontes de informação distintas. Os resultados poderiam indicar divergências estatisticamente significativas na linguagem entre as fontes.

Web crawlers e *web scrapers* são programas que automatizam a busca e a seleção de informações armazenadas em *sites* para processamento posterior pelo emprego de alguma técnica analítica. Funcionam por meio do uso de rastreadores que seguem *links* de hipertexto de um *site* para o outro e atualizam a cadeia de referências entre *sites*, de modo a gerarem uma rede conectada que viabiliza a busca recursiva de conteúdo.

Entre as modalidades analíticas sugeridas para o processamento de frações de interesse da atividade de Inteligência, estão as análises léxica, de conexões em rede social (*Social network analysis*), de sentimentos, de georreferenciamento e geoinferência, além da técnica de detecção de eventos.

A análise léxica efetua-se por testes estatísticos para contar a frequência de palavras, a distância entre palavras e outras características para detectar estruturas e padrões em frações do texto. É usada com mais frequência para determinar empiricamente a que se refere uma coleção de textos, por meio de palavras visivelmente super e sub-representadas, e as conexões entre as palavras presentes na coleção de textos. A modalidade pode ser empregada para classificar documentos e agregar informações semelhantes.

A *Social network analysis* (SNA) envolve a identificação e a visualização de estruturas sociais. Baseia-se no trabalho multidisciplinar em Psicologia, Sociologia e Teoria dos Grafos da Matemática. Essa modalidade analítica busca revelar a natureza, a intensidade e a frequência dos relacionamentos estabelecidos, na pressuposição de que os laços sociais influenciam as crenças, os comportamentos e as experiências das pessoas. Por meio de algoritmos de aferição e mapeamento desses relacionamentos, a SNA tenta explicar e prever o comportamento dos indivíduos dentro da rede.

A análise de sentimento abarca a identificação, a extração e a enumeração da atitude do usuário em relação às informações que são por si fornecidas em um texto de formato livre. Com essa técnica, é possível, com base em uma coleção de textos (comentários ou postagens de rede social, por exemplo), revelar a predominância de sentimentos, positivos, negativos ou neutros, em relação a algum acontecimento, ou mesmo revelar concordância ou divergência de pontos de vista em relação a algum assunto.

Georreferenciamento e geoinferência são dois métodos utilizados para determinar a origem geográfica de uma mensagem postada na plataforma de rede social. O georreferenciamento vale-se do registro de coordenadas geográficas e é altamente

preciso; no entanto, caso o usuário tenha desativado o recurso de localização geográfica instalado no dispositivo, é possível, por geoinferência, considerar os metadados capturados para fazer inferências sobre a localização geográfica das postagens com consideráveis níveis de precisão.

A técnica de detecção de eventos baseia-se no processamento das mensagens postadas sobre eventos atuais, a fim de proceder com uma classificação conforme o tipo de evento (especificado ou não-especificado), a tarefa de detecção (retrospectiva ou detecção de novo evento) e o método de detecção (supervisionado ou não-supervisionado). A capacidade multimídia (imagens, áudios e vídeos) das plataformas de rede social favorece a consciência situacional e satisfaz a credibilidade do conhecimento produzido acerca do evento, de modo que os usuários da rede social passam a figurar também fontes de informação (*crowd-sourced information*).

A depender das informações disponíveis acerca de um evento de interesse, a detecção pode ser trabalhada para traçar perspectivas sobre um evento específico ou sobre uma categoria de eventos. Quando não há informações suficientes sobre um evento, toma-se por base marcos temporais para detectar fluxos de informações que conduzam à aferição de condições e de tendências sobre a ocorrência do evento.

Essas técnicas requerem o monitoramento de postagens em perfis de interesse na rede social, o agrupamento de fatos que identificam tendências e a classificação dos eventos em diferentes categorias. Uma vez estabelecidas as categorias, é possível a aplicação de técnicas para recuperação e extração de informações, tais como filtragem de frações significativas, elaboração de consultas personalizadas, agrupamentos (*clustering*) de categorias e agregação de dados.

Após a captura, a seleção e a organização dos dados, conforme a técnica mais conveniente para o tipo de dado tratado e para a modalidade analítica empregada, cabe automatizar o processamento com técnicas da Ciências de Dados que envolvem o uso de mecanismos de Inteligência Artificial. Entre eles estão o Processamento de Linguagem Natural, o Aprendizado de Máquina, o Aprendizado Profundo e as Redes Neurais Artificiais.

O Processamento de Linguagem Natural (*Natural Language Processing* – NLP) envolve analisar, compreender e gerar respostas a fim de permitir a interação com sistemas computadorizados que recebem extratos de expressão da linguagem humana como insumo para processamento e, por vezes, para prover uma resposta ao emissor da mensagem de entrada. A NLP favorece a síntese e a tradução de textos, além de ser a tecnologia pilar dos sistemas de reconhecimento de voz, a exemplo dos *chatbots*. Para textos extraídos de postagens ou comentários em redes sociais, a NLP normalmente é empregada para analisar semanticamente frases emitidas sobre entidades (pessoas, lugares, organizações), conceitos (indicam uma ideia específica), temas (arranjos de conceitos) ou mesmo sentimentos.

O Aprendizado de Máquina (*Machine Learning* – ML) envolve a utilização de algoritmos para extrair informações de dados brutos e representá-los por meio de um modelo matemático, que se ajusta a novas circunstâncias e que detecta e extrapola padrões. ML se propõe a identificar padrões para gerar hipóteses a partir de dados com mínima intervenção humana. O computador aprende por meio de exemplos em abordagem supervisionada ou não-supervisionada. No Aprendizado supervisionado, o algoritmo de aprendizado (indutor) recebe uma massa de dados de treinamento para conceber um classificador que possa determinar quais dados exemplificativos fornecidos seriam classificados sob um rótulo. No Aprendizado não-supervisionado, o indutor verifica a massa de dados exemplificativa fornecida sem um rótulo especificado e se propõe a determinar um possível agrupamento (*clustering*).

Aprendizado Profundo (*Deep Learning*) é uma subárea da ML que emprega algoritmos para processar e simular o processamento feito pelo cérebro humano. Usa camadas com fórmulas matemáticas, que simulam o funcionamento de um neurônio, para processar dados, compreender a fala humana e reconhecer objetos visualmente. A informação é passada através de cada camada, de modo que o resultado do processamento da camada anterior seja o parâmetro de entrada para a próxima camada. Baseia-se no conceito de Redes Neurais Artificiais.

Redes Neurais Artificiais (*Deep neural networks* – DNNs) são redes multicamadas que permitem a captura e a mineração de maiores volumes de dados, incluindo dados não-estruturados. DNNs visam a reconhecer padrões ocultos e correlações em dados brutos, para agrupá-los e classificá-los de forma contínua para a melhoria da capacidade cognitiva do sistema específico. Em virtude dessa aptidão das DNNs, o Aprendizado Profundo tornou-se o responsável por avanços em visão computacional, reconhecimento de fala, NLP e reconhecimento de áudio. DNNs podem se classificar em convolucional ou recorrente.

Redes Neurais Convolucionais contêm camadas de entradas, de convolução, de agrupamento, de saída e são conectadas a fim de cumprir com um propósito específico, tal como síntese ou conexão. Além da classificação de imagens e da percepção de objetos, as redes neurais convolucionais aplicam-se em áreas como previsão de cenários e NLP.

Redes Neurais Recorrentes se valem de informações sequenciais, a exemplo de uma sentença ou do registro de data e hora coletado por um sensor. As entradas de uma rede neural recorrente são interdependentes, e os resultados para cada elemento dependem da computação dos elementos precedentes. Esse tipo de rede neural é normalmente utilizado na previsão e na aplicação de séries temporais, na análise de sentimento e em outras aplicações baseada em dados textuais.

As abordagens analíticas de frações coletadas mais promissoras para a satisfação das atribuições da atividade

de Inteligência, ao se considerar a análise semântica dos dados que são coletados em plataformas de rede social, em maiores variedade e volume, ou seja, preponderantemente aqueles no formato textual e imagens com texto embutido, são a detecção de eventos e a conjunção de Processamento de Linguagem Natural associado ao Aprendizado de Máquina via Redes Neurais Artificiais.

Abordagens práticas de inteligência de dados de redes sociais digitais

As técnicas da Ciência de Dados ganham importância no desenvolvimento de soluções das áreas de Segurança, de Defesa e de Inteligência, em especial pelos amplos volume e variedade dos dados coletados de plataformas de redes sociais. Ao explorar a miríade de informações, é possível encontrar menções a atividades de grupos que atentem contra a ordem pública e contra a segurança da sociedade e do Estado. Para cumprir com a oportunidade na difusão do conhecimento, é necessário se pensar na automação da extração das informações, de modo que se detecte, com a máxima antecipação possível, uma ameaça sinalizada por comentário ou texto publicado.

Entre as aplicações de inteligência de dados de redes sociais está a detecção da localização geográfica dos usuários por meio da publicação no perfil. O resultado pode ser útil para revelar eventos ou atividades ocorridas em locais específicos. Por exemplo, potenciais planos de terroristas podem considerar como alvos áreas geográficas específicas. Extrair a localização dos usuários com base em postagens de mídia social ou em metadados de rede social também pode ajudar, uma vez que nem todos os usuários declaram ou alguns conseguem forjar sua localização no perfil da plataforma de rede social.

Ao se considerar os metadados sobre a localização dos usuários, pode-se treinar um classificador que preveja a localização de qualquer usuário e, portanto, viabilize geoinferência. O classificador pode captar diferenças sutis na linguagem (dialetos) e nos tipos de entidades mencionadas. A localização do usuário pode ser estimada ao se submeter ao classificador um conjunto de mensagens perturbadoras publicadas, caracterizadas como discurso de ódio ou como incitação a crimes. Ainda que o usuário tente forjar uma localização, ao declará-la explicitamente no perfil pessoal, o classificador pode ser usado para detectar declarações falsas, uma vez que se baseará nos metadados da publicação.

Hecht *et alii* (2011) conduziram experimentos de Aprendizado de Máquina para identificar a localização de um usuário ao verificar apenas sobre o que o usuário publicou no Twitter, ou seja, em um esforço de geoinferência. Os cientistas demonstraram que a localização geográfica do usuário pode ser determinada automaticamente com precisão razoável, ao indicar que os usuários revelam implicitamente informações de localização, mesmo que inconscientemente. Essa constatação sugere implicações éticas e legais para serviços baseados em localização e pode

levantar questões acerca da privacidade dos usuários do serviço.

Outra abordagem refere-se à análise semântica textual. A análise de um texto extraído de uma fonte pode revelar tendências e vieses propalados pelo veículo de informações. Kaati *et alii* (2016) aplicaram a técnica a sítios da Internet de mídia sueca com expressões de repúdio em desfavor ao acolhimento de imigrantes. Procedeu-se a um filtro para detecção de narrativas que continham estereótipos xenófobos e conspiratórios. Por meio da ferramenta de análise de texto *Linguistic Inquiry and Word Count* (LIWC[22]), com base no léxico sueco presente em um conjunto de dicionários, processaram narrativas xenófobas capturadas em *sites* tradicionais e em alternativos. Os resultados indicaram uma divergência estatisticamente significativa na linguagem entre os *sites* de mídia convencional e os *sites* alternativos críticos.

A detecção de emoções e a análise de sentimentos, ambos os recursos de Processamento de Linguagem Natural, aplicadas a publicações em redes sociais representam técnicas de interesse da Inteligência. A detecção de reações de consternação, raiva ou desapontamento é de particular interesse. Classificadores de emoção (incluindo raiva e tristeza) foram testados em dados de *blogs* (GHAZI *et alii,* 2010), e em dados de plataforma unificada de hospedagem de diários e de periódicos divulgados na Internet (KESHTKAR; INKPEN, 2009).

Mensagens de ódio a países podem ser de autoria de potenciais perpetradores de ameaças terroristas. Combinada com a detecção de tópicos, a detecção de sentimentos pode levar a uma indicação mais precisa das ameaças em potencial. A detecção de discursos de desânimo ou indiferença em publicações pode sugerir a autoria por pessoas que nutrem tendências suicidas ou jovens que não têm um senso de pertencimento e podem ser tentados a aderir a atividades extremistas ou terroristas. Essa análise pode ser combinada com *Social Network Analysis,* e um usuário pode ser sinalizado como potencialmente perigoso quando houver registros de relacionamento com indivíduos suspeitos já identificados.

Técnicas de análise de sentimento também podem ser utilizadas para detectar opiniões sobre eventos sociais e políticos. Colbaugh e Glass (2010) desenvolveram estudo de caso que envolveu a estimativa do sentimento público dos indonésios em relação aos atentados a hotéis em julho de 2009 em Jacarta, capital da Indonésia.

Nos últimos anos, organizações terroristas, como o Estado Islâmico no Iraque e *al-Sham* (ISIS), aumentaram suas interações em plataformas de redes sociais com vistas a recrutar e a promover radicalização de cidadãos de países ocidentais. Rowe e Saif (2016) examinaram os hábitos em plataforma de rede social de usuários que se radicalizaram. Os pesquisadores definiram um conjunto de contas na plataforma Twitter associadas a publicações referentes ao conflito na Síria.

22 LIWC é um *software* de análise de texto que organiza palavras em categorias derivadas de gramática e de psicologia.

Então, ao contemplar os seguidores dessas contas, produziram uma lista de 154 mil usuários que viviam na Europa. Dessa lista, coletaram 104 milhões de postagens com o objetivo de examinar o comportamento do usuário antes e depois da radicalização.

Rowe e Saif (2016) avaliaram os conteúdos compartilhados e os padrões linguísticos dos usuários, a fim de determinar se eles eram favoráveis ou contrários à ideologia difundida pelo ISIS. Para a análise linguística, considerou-se um léxico com termos de exaltação favorável aos ideais do ISIS. Os resultados da análise demonstraram que o uso de termos favoráveis ao ISIS aumenta dramaticamente após a radicalização.

Abordagem pragmática complementar tangencia a consecução da consciência situacional acerca de evento ou conjuntura. Yin *et alii* (2015) implementaram um sistema que extraiu informações de publicações de usuários na plataforma Twitter (*tweets*) durante diversos desastres e crises ocorridas em 2010 e 2011 na Austrália e na Nova Zelândia. Os dados continham 66 milhões *tweets* de aproximadamente 2,51 milhões de perfis distintos que cobriam uma variedade de desastres e incidentes de segurança, entre eles o ciclone tropical Ului (março de 2010), as tempestades em Brisbane (junho de 2010), o atentado perpetrado por um atirador em Melbourne (junho de 2010), os terremotos de Christchurch (setembro de 2010 e fevereiro de 2011), o incidente que envolveu aviões Qantas A380 (novembro de 2010), as inundações de Brisbane (janeiro de 2011) e o ciclone tropical Yasi (fevereiro de 2011). Por meio de recursos lexicais, o sistema buscou identificar denominadores comuns para os incidentes inesperados, a fim de estabelecer uma classificação que subsidiasse a avaliação de impacto do incidente.

Nunes *et alii* (2016) desenvolveram um sistema operacional capaz de detectar ameaças cibernéticas em sites de mídia social hospedados em *Darknet*[23]. O sistema implementou um *web crawler*, um analisador sintático e um classificador. Os pesquisadores desenvolveram um modelo preditivo de Aprendizagem supervisionada que considera algoritmos para tratar problemas de classificação, a fim de determinar categorias, e de regressão, para definir um valor numérico indicativo de escore. Essa abordagem alcançou uma pontuação de confirmação na ordem de 0,92 em plataformas de transações ilegais e 0,8 em fóruns de discussão relacionados a ações para comprometer sistemas (*hacking*).

Thorleuchter e Van Den Poel (2013) investigaram a proteção de projetos de Pesquisa e Desenvolvimento (P&D) contra a espionagem. Os pesquisadores desenvolveram um sistema de identificação de padrões textuais semânticos para representar tecnologias e correspondentes campos de aplicação tidos como de alta relevância para a estratégia de uma organização.

23 *Darknet* é o termo de referência ao ambiente da Internet, acessível por *software* específico, onde as comunicações prezam pela manutenção do anonimato, em especial por oferecer plataformas de transação de produtos ilegais, como armas, drogas e órgãos humanos. Entre as plataformas, estão Empire Market, Icarus Market, Dark0de Reborn, Deep Sea Market, The Versus Project, Canada HQ, Monopoly Market, ToRReZ Market e Hydra Market.

Nessa pesquisa, os padrões semânticos foram usados para estimar os custos organizacionais provocados pelo vazamento de informações para cada projeto. Uma abordagem de mineração de dados em fontes de conhecimento científico na Internet foi conduzida para identificar a distribuição do conhecimento em tecnologias estratégicas em escala mundial. Essa informação foi usada para estimar a probabilidade de ocorrer um vazamento de informações. Uma metodologia de avaliação de risco veio calcular o risco de vazamento de informações para cada projeto.

Os pesquisadores concluem a pesquisa com um estudo de caso que se propõe a estimar o risco de vazamento de projetos de pesquisa de tecnologias da área de Defesa, que é de particular interesse para a espionagem. No geral, a metodologia proposta se mostrou exitosa no cálculo do risco de vazamento de informações dos projetos de Defesa. Logo, demonstraram que é possível estabelecer o gerenciamento de risco de espionagem para auxiliar na proteção de uma organização que considera estratégicos os esforços em P&D.

Em abordagem alternativa, Zhou *et alii* (2011) associaram imagens e descrições feitas sobre elas em forma de palavras de marcação (*tags*) para definir um *framework*[24] probabilístico de recomendação de *tags* complementares que possam rotular imagens sugeridas como de interesse para alguma necessidade de acompanhamento. Imagens e legendas de texto associadas podem ser coletadas em postagens de redes sociais, onde é popular a marcação colaborativa de imagens em forma de comentários ou por meio do uso de recurso que adiciona textos às imagens publicadas.

Considerações finais

O presente artigo apresentou aplicações associadas de técnicas da Ciência de Dados que demandam o aperfeiçoamento constante para a produção de conhecimentos de Inteligência, a fim de dar conta das transformações pelas quais a sociedade passou nos últimos vinte anos. As abordagens práticas de inteligência de dados de redes sociais digitais buscaram apresentar a viabilidade do emprego de técnicas da Ciência de Dados para solucionar demandas comuns no cotidiano da atividade de Inteligência. Entre essas técnicas, estão mineração de dados, geoinferência, Aprendizado de Máquina, Processamento de Linguagem Natural, análise textual, análise de sentimentos, *Social Network Analysis*, identificação de padrões semânticos e construção de um sistema de recomendação.

As pesquisas científicas relatadas apresentaram evidências da aplicação das técnicas referenciadas em diversos objetos e situações de interesse da atividade Inteligência. A prática recorrente na produção de conhecimento de inteligência nos órgãos congêneres aponta para a adoção das técnicas não só para

24 Um *framework* define uma arquitetura para uma família de subsistemas e oferece os construtores básicos para criá-los. Também são explicitados os lugares ou pontos de extensão (*hot-spots*) nos quais adaptações do código devem ser feitas para funcionamento específico de determinados módulos (BUSCHMANN *et alii*, 1996).

automatizar grande parte do trabalho do profissional de Inteligência, mas, sobretudo, para permitir analisar redes sociais digitais, o que seria impossível sem essas técnicas.

A evolução de tecnologias de comunicação móvel e de computação pervasiva favorecem a intensidade na adoção e no uso de plataformas de redes sociais. Essa constatação sugere que novas pesquisas devem se esforçar no desenvolvimento de novos algoritmos e métodos baseados em Ciência de Dados e Inteligência Artificial.

A Ciência de Dados, por ter como finalidade produzir conhecimento (a partir de dados) para a tomada de decisão, guarda uma identidade com a ideia de *actionable intelligence*, de sorte que, como conjunto de métodos e técnicas, é essencial para analisar redes sociais digitais. A viabilidade da análise de dados de redes sociais é potencializada pelo emprego de Aprendizado de Máquina e de Processamento de Linguagem Natural contidos na Ciência de Dados. A Ciência de Dados favorece a geração de *actionable intelligence* como nenhum outro ferramental. Logo, é necessária a criação de condições para o aprimoramento do uso sistemático de técnicas da Ciência de Dados para que o conhecimento de Inteligência trate de todos os temas relevantes e para que as análises produzidas sejam cada vez mais tidas como imprescindíveis ao Processo Decisório Nacional.

Referências

AMBROS, Christiano; LODETTI, Daniel. Vieses cognitivos na atividade de Inteligência: conceitos, categorias e métodos de mitigação. *Revista Brasileira de Inteligência*. Brasília, n. 14, p. 9-34, dez. 2019.

ANANTHANARAYANAN, Rajagopal; ESSER, Steven; MODHA, Dharmendra; NDIRANGO, Anthony; SHERBONDY, Anthony; SINGH, Raghavendra. Cognitive Computing. *Communications of the ACM*, v. 54, p. 62-71, 2011. Disponível em: http://cacm.acm.org/magazines/2011/8/114944-cognitive-computing/fulltext. Acesso em: 01 out. 2021.

BRASIL, *Lei nº 9.883*, de 7 de dezembro de 1999. Institui o Sistema Brasileiro de Inteligência, cria a Agência Brasileira de Inteligência - ABIN, e dá outras providências. Brasília, DF: Presidência da República, 1999. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/L9883.htm. Acesso em: 01 out. 2021.

_____. *Decreto nº 4.376*, de 13 de setembro de 2002. Dispõe sobre a organização e o funcionamento do Sistema Brasileiro de Inteligência, instituído pela Lei nº 9.883, de 7 de dezembro de 1999, e dá outras providências. Brasília, DF: Presidência da República, 2002. Disponível em: http://www.planalto.gov.br/ccivil_03/decreto/2002/d4376.htm. Acesso em: 01 out. 2021.

_____. *Decreto nº 8.793*, de 29 de junho de 2016. Fixa a Política Nacional de Inteligência. Brasília, DF: Presidência da República, [2016a]. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/D8793.htm. Acesso em: 01 out. 2021.

_____ . Gabinete de Segurança Institucional. Agência Brasileira de Inteligência. *Doutrina Nacional de Inteligência*: fundamentos doutrinários fundamentos doutrinários. Brasília: ABIN, 2016 Disponível em: https://www.gov.br/abin/pt-br/centrais-de-conteudo/publicacoes/Col3v58.pdf. Acesso em: 01 out. 2021.

_____. *Decreto nº 10.445*, de 30 de julho de 2020. Aprova a Estrutura Regimental e o Quadro Demonstrativo dos Cargos em Comissão e das Funções de Confiança da Agência Brasileira de Inteligência e remaneja e transforma cargos em comissão e funções de confiança. Brasília, DF: Presidência da República, 2020. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2020/decreto/ D10445.htm. Acesso em: 1º out. 2021.

BURLINGAME, Noreen; NIELSEN, Lars. *A Simple Introduction to Data Science*. Wickford, New Street Communications, 2012.

BUSCHMANN, Frank; MEUNIER, Régine; ROHNERT, Hans; SOMMERLAD, Peter; STAHL, Michael. *Pattern-Oriented Software Architecture* - A System of Patterns. New York-NY: John Wiley and Sons, 1996.

CLARKE, Roger. Information technology and dataveillance. *Communications of the ACM*. v. 31, n. 5, p. 498-512, 1988. Disponível em: http://www.rogerclarke.com/DV/CACM88.html. Acesso em: 01 out. 2021.

COLBAUGH, Richard; GLASS, Kristin. Estimating sentiment orientation in social media for intelligence monitoring and analysis. *IEEE International Conference on Intelligence and Security Informatics (ISI)*, p. 135–137, 2010. Disponível em: https://www.scss.tcd.ie/Khurshid.Ahmad/Research/Sentiments/K_Teams_Buchraest/05484760.pdf. Acesso em: 01 out. 2021.

GANTZ, John; REINSEL, David. *Extracting Value from Chaos*, Framingham: International Data Corporation. 2011. Disponível em: https://www.whizpr.be/upload/medialab/21/company/IDC_1142.pdf . Acesso em: 01 out. 2021.

GHAZI, Diman, INKPEN, Diana; SZPAKOWICZ, Stan. *Hierarchical versus flat classification of emotions in text. Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*. Association for Computational Linguistics, p. 140–146, 2010. Disponível em: https://dl.acm.org/doi/pdf/10.5555/1860631.1860648. Acesso em: 01 out. 2021.

GILLILAND, Anne J. *Introduction to Metadata*. 3ª ed. Los Angeles, Getty Research Institute, 2016. Disponível em: http://www.getty.edu/publications/intrometadata/. Acesso em: 01 out. 2021.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. *Data Mining*: Concepts and Techniques. 3ª ed. Waltham, Morgan Kaufmann Publishers, 2012.

HAYASHI, Chikio. What is Data Science? Fundamental Concepts and a Heuristic *Example. In: Data Science, Classification, and Related Methods*. Studies in Classification, Data Analysis, and Knowledge Organization. Tokyo, Springer, 1998.

HECHT, Brent; HONG, Lichan; SUH, Bongwon; CHI, Ed H. Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11). Vancouver: Canada. ACM. p. 237–246, 2011. Disponível em: https://www-users.cs.umn.edu/~bhecht/publications/bhecht_chi2011_location.pdf. Acesso em: 19 out. 2021.

KAATI, Lisa; SHRESTHA, Amendra; COHEN, Katie; LINDQUIST; Sinna. *Automatic detection of xenophobic narratives: A case study on swedish alternative media. IEEE Conference on Intelligence and Security Informatics (ISI)*. Tucson: USA. IEEE. 2016. Disponível em: https://www.foi.se/download/18.7fd35d7f166c56ebe0bffd8/1542623691578/Automatic-detection-xenophopic_FOI-S--5655--SE.pdf. Acesso em: 01 out. 2021.

KESHTKAR, Fazel; INKPEN, Diana. *Using Sentiment Orientation Features for Mood*

Classification in blogs. IEEE International Conference on Natural Language Processing and Knowledge Engineering. Dalian: China. IEEE. 2009. Disponível em: https://www.researchgate.net/publication/224076625_Using_sentiment_orientation_features_for_mood_classification_in_blogs. Acesso em: 01 out. 2021.

KELLEHER, John D; TIERNEY, Brendan. *Data Science*, 1ª ed. Cambridge, MIT Press, 2018.

LOWENTHAL, Mark. *Intelligence: from secrets to policy*. 8ª ed. Washington-DC, CQ Press, 2019.

NUNES, Eric; DIAB, Ahmad; GUNN, Andrew; MARIN, Ericsson; MISHRA, Vineet; PALIATH, Vivin; ROBERTSON, John; SHAKARIAN, Jana; THART, Amanda; SHAKARIAN, Paulo. *Darknet and Deepnet mining for proactive cybersecurity threat intelligence*. 2016. Disponível em: https://arxiv.org/abs/1607. 08583. Acesso em: 01 out. 2021.

O'NEIL, Cathy. *Weapons of Math Destruction: how big data increases inequality and threatens democracy*. New York, Crown Publishers, 2016.

ROWE, Matthew; SAIF, Hassan. *Mining pro-isis radicalisation signals from social media users. Proceedings of the Tenth International AAAI Conference on Web and Social Media* (ICWSM 2016). Cologny: Germany. AAAI. p. 329–338, 2016. Disponível em: http://oro.open.ac.uk/48477/. Acesso em: 01 out. 2021.

RUSSELL, Stuart Jonathan; NORVIG, Peter. *Artificial Intelligence*: A Modern Approach. 4ª ed. Global edition. Hoboken, Pearson, 2021.

SHERMAN, Chris; PRICE, Gary. *The invisible web: uncovering information sources:* search engines can't see. 7ª ed. Medford, CyberAge Books, Information Today, Inc., 2001.

THORLEUCHTER, Dirk; VAN DEN POEL; Dirk. Protecting research and technology from espionage. *Expert Systems with Applications.* Elsevier. v. 40, issue 9, p. 3432-3440. 2013. Disponível em: http://wps-feb.ugent.be/Papers/wp_12_824.pdf Acesso em: 1º out. 2021.

WANG, Yingxu; ZHANG, Du; LATOMBE, Jean-Claude; KINSNER, Witold. Advances in the Fields of Cognitive Informatics and Cognitive Computing. *In: Advances in Cognitive Informatics and Cognitive Computing*. Stanford, Springer, 2010.

YIN, Jie; LAMPERT, Andrew; CAMERON, Mark; ROBINSON, Bella; POWER, Robert. Using social media to enhance emergency situation awareness. *International Joint Conference on Artificial Intelligence* Buenos Aires: Argentina. IEEE. v. 27, p. 52–59, 2015. Disponível em: https://www.researchgate.net/publication/280829031_Using_Social_Media_to_Enhance_Emergency_Situation_Awareness_Extended_Abstract. Acesso em: 01 out. 2021.

ZHOU, Ning; CHEUNG, William K.; QIU, Guoping; XUE, Xiangyang. A hybrid probabilistic

model for unified collaborative and content-based image tagging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE. v. 33, p. 1281–1294, 2011. Disponível em: https://www.researchgate.net/publication/224196190_A_Hybrid_Probabilistic_Model_for_Unified_Collaborative_and_Content-Based_Image_Tagging. Acesso em: out. 2021.

Artigo recebido em 31 ago. 2021
Aprovado em 13 set. 2021