Desinformação no Meio de Segurança

Luis João Victoria Vinicius

Agenda

Tempo	Conteúdo
0–15 min	**Fundamentos teóricos** (taxonomias, modelos formais, arquiteturas de detecção, complexidade computacional)
15–30 min	**Estudo de caso:** Análise de Benevenuto & Melo (CACM 2024) sobre campanhas de desinformação via WhatsApp/Telegram

Taxonomia Formal de Desinformação

- **Desinformação** $D = \{m \in M : intent(m) = malicious \land veracity(m) = false\}$
- **Misinformation** $Mi = \{m \in M : intent(m) = benign \land veracity(m) = false\}$
- **Mal-information** $Ma = \{m \in M : intent(m) = malicious \land veracity(m) = true \land context(m) = manipulated\}$
- Onde M é o conjunto de todas as mensagens no espaço informacional

Teoria dos Grafos: Propagação em Redes Sociais

- Grafo direcionado $G = (V, E)$ onde V são usuários e E são conexões
- **Modelo de Cascata Independente:** $P(v \text{ ativa } u) = p_{v,u}$
- **Modelo de Limiar Linear:** $f_v(S) = \sum_{u \in S} b_{v,u}$, ativa se $f_v(S) \geq \theta_v$
- **Centralidade de intermediação:** $BC(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$
- Detecção de bots via análise de propriedades topológicas não-humanas

Arquitetura de Inteligência de Ameaças

- **Framework MITRE ATT&CK** para desinformação
- **Modelo Diamante:** Adversário - Infraestrutura - Capacidade - Vítima
- **Cadeia Cibernética de Ataque adaptada:** Reconhecimento → Armamento → Entrega → Exploração → Instalação → C2 → Ações
- Integração com ferramentas de OSINT (Maltego, Shodan, ThreatConnect)

Taxonomia de Táticas e Técnicas Adversariais

Tática	Técnicas Específicas
Armamento	Fabricação de deepfakes, manipulação de contexto, criação de narrativas falsas
Entrega	Spam coordenado, fazendas de bots, contas falsas, astro-turfing
Persistência	Disseminação multiplataforma, câmaras de eco, bolhas de filtro
Comando e Controle	Coordenação via canais privados, amplificação de sinais, manipulação de tendências
Evasão	Migração entre plataformas, mutação de conteúdo, ofuscação semântica

Complexidade Computacional: Detecção vs. Adversários

- **Problema de Detecção:** NP-difícil para grafos gerais
- **Teoria dos Jogos:** $\max_d \min_a U(d, a)$ onde d são estratégias de detecção, a são estratégias adversariais
- **Equilíbrio de Nash:** (d^*, a^*) tal que nenhum jogador melhora unilateralmente
- **Complexidade de Amostragem:** $O(n \log n)$ para detecção com alta probabilidade

Pipeline de PLN para Classificação

Algoritmo de Classificação de Desinformação:

1. **Entrada:** Texto t, modelo pré-treinado M
2. $tokens \leftarrow tokenizar(t)$
3. $incorporações \leftarrow M.codificar(tokens)$
4. $características \leftarrow [características\_linguísticas(t), características\_rede(t)]$
5. $combinado \leftarrow concatenar(incorporações, características)$
6. $logits \leftarrow classificador(combinado)$
7. $prob \leftarrow softmax(logits)$
8. **retorna** $argmax(prob)$

Métricas: Precisão, Revocação, F1-score, AUC-ROC, Coeficiente de Correlação de Matthews

Modelos Multimodais: Fusão de Modalidades

- **Fusão Precoce:** $f(x_t, x_v, x_a) = MLP(concat(x_t, x_v, x_a))$
- **Fusão Tardia:** $f(x_t, x_v, x_a) = \sigma(\alpha f_t(x_t) + \beta f_v(x_v) + \gamma f_a(x_a))$
- **Mecanismo de Atenção:** $Atenção(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$
- **Transformers Intermodais** para detecção de inconsistências semânticas

Criptografia e Verificação de Integridade

- **Árvores de Merkle** para marcação temporal de conteúdo: $h = H(H(h_1||h_2)||H(h_3||h_4))$
- **Provas de Conhecimento Zero** para verificação sem exposição de dados sensíveis
- **Assinaturas Digitais:** RSA, ECDSA para autenticação de fonte
- **Proveniência baseada em Blockchain** com contratos inteligentes para rastreabilidade

Métricas de Avaliação e Benchmarks

- **Conjuntos de Dados:** FakeNewsNet, LIAR, CoAID, CONSTRAINT-2021
- **Métricas além da acurácia:**
  - Justiça: $DP = P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1)$
  - Robustez: acurácia adversarial sob ataques ℓ_∞
  - Explicabilidade: valores LIME, SHAP
- **Frameworks de Avaliação:** FEVER, CheckThat!, CLEF

Casos de Uso Globais em Desinformação

- **Agência de Pesquisa da Internet (IRA) - Rússia:** Interferência nas eleições americanas 2016/2020
  - 126 milhões de usuários alcançados no Facebook, 20 milhões no Instagram
  - Coordenação multiplataforma (Facebook, Twitter, Instagram, YouTube)
  - Investimento de $1.25 milhão/mês, 3.393 anúncios pagos
  - Técnicas: personas falsas, eventos organizados, amplificação artificial
- **SolarWinds/Sunburst (APT29):** Ataques à cadeia de suprimentos + desinformação
  - 18.000 organizações comprometidas globalmente (Microsoft, FireEye, Cisco)
  - Desinformação sobre atribuição, minimização do impacto real
  - Operação orquestrada por 18 meses sem detecção

Casos de Uso Globais em Desinformação (cont.)

- **Campanhas Anti-Vacina e COVID-19:** Desinformação em saúde pública
  - "Dozen Disinformers": 12 indivíduos geraram 65% da desinformação anti-vacina
  - Movimento coordenado internacional com receita estimada em $36 milhões
  - Técnicas: cherry-picking de estudos, apelos emocionais, teorias conspiratórias
- **Guerra na Ucrânia (2022-presente):** Operações de informação em tempo real
  - Vídeos falsificados profundos de autoridades ucranianas (Zelensky)
  - Narrativas falsas: "laboratórios biológicos", "nazistas", "genocídio"
  - Coordenação RT, Sputnik: 7 milhões de interações/dia pré-guerra
  - Técnicas: bots multiplataforma, influenciadores pagos, deepfakes

Metodologia Científica do Estudo

Artigo: "Misinformation Campaigns Through WhatsApp and Telegram in Presidential Elections in Brazil"(CACM 2024)

- **Conjunto de Dados:** 1.2M mensagens WhatsApp + 3M tweets (2018-2022)
- **Rotulação:** Colaboração com Lupa e Aos Fatos (verificadores certificados)
- **Metodologia:** Abordagem de métodos mistos combinando análise quantitativa e qualitativa
- **Aprovação do CEP** e considerações éticas para dados sensíveis

Arquitetura Técnica de Coleta

- **WhatsApp:** Selenium WebDriver + engenharia reversa da API WhatsApp Web
- **Telegram:** API Bot Oficial + raspagem de canais públicos
- **Twitter:** API de Pesquisa Acadêmica v2 com limitação de taxa
- **Pré-processamento:** Normalização UTF-8, desduplicação via MinHash LSH
- **Armazenamento:** MongoDB para dados não-estruturados, PostgreSQL para metadados

Algoritmos de Detecção Implementados

- **Baseado em BERT:** Ajuste fino do BERTimbau com perda de entropia cruzada
- **Análise de Redes:** PageRank modificado para detecção de influenciadores
- **Análise Temporal:** Janela deslizante + detecção de rajadas usando aproximação de Poisson
- **Detecção de Bots:** Características engenheiradas: frequência de tweets, entropia temporal, razão seguidor-seguindo
- **Métodos de Conjunto:** Floresta Aleatória + Gradient Boosting com ponderação de votos

Resultados Quantitativos do Artigo

Métrica	**Resultado**
F1-score (BERTimbau)	0.92 ($\pm$0.03)
Acurácia Detecção de Bots	0.89 usando Botometer-PT
Precisão@10 (classificação)	0.87 para conteúdo viral falso
Correlação Temporal	$\rho = 0.74$ entre eventos políticos e picos de desinformação

Significância Estatística: Todos os resultados com $p < 0.01$ (teste t de Student)

Análise de Redes: Descobertas Topológicas

- **Propriedade de Mundo Pequeno:** Diâmetro médio $d = 6.2$ (vs. $d = 4.1$ para redes orgânicas)
- **Distribuição Lei de Potência:** $P(k) \sim k^{-\gamma}$ com $\gamma = 2.8$ (livre de escala)
- **Coeficiente de Agrupamento:** $C = 0.67$ (vs. $C = 0.15$ esperado para grafos aleatórios)
- **Detecção de Comunidades:** Algoritmo de Louvain identificou 847 comunidades distintas
- **Assortatividade:** $r = -0.23$ (mistura desassortativa, hubs conectam com nós pequenos)

Comportamento Inautêntico Coordenado (CIC)

- **Definição formal:** Conjunto de contas S onde $\forall s_i, s_j \in S$:
  - $\mathit{sobreposição\_temporal}(s_i, s_j) > \theta_t$
  - $\mathit{similaridade\_conteúdo}(s_i, s_j) > \theta_c$
  - $\mathit{proximidade\_rede}(s_i, s_j) > \theta_n$
- **Detecção:** Algoritmo baseado em agrupamento de grafos + sincronia temporal
- **Resultados:** 1,247 grupos identificados com > 5 contas cada

Impacto Quantificado na Eleição

- **Estimativa de Alcance:** Modelo epidemiológico SIR adaptado
- **Métricas de Exposição:** $E = \sum_{i=1}^{n} v_i \cdot l_i \cdot T_i$ (visualizações $\times$ influência $\times$ tempo)
- **Inferência Causal:** Design diferença-em-diferenças comparando regiões
- **Tamanho do Efeito:** d de Cohen $= 0.31$ para mudança de intenção de voto (efeito médio)
- **Intervalos de Confiança:** [0.18, 0.44] com 95% de confiança

Limitações e Trabalhos Futuros

- **Limitação 1:** Viés de seleção devido à disponibilidade limitada de dados do WhatsApp
- **Limitação 2:** Escopo temporal limitado (não captura evolução de longo prazo)
- **Trabalho Futuro 1:** Aprendizado federado para análise preservando privacidade
- **Trabalho Futuro 2:** Detecção multimodal de deepfake usando GANs adversariais
- **Trabalho Futuro 3:** Detecção em tempo real com processamento de fluxo (Apache Kafka + Storm)

Conclusões Técnicas

- Algoritmos de AM alcançam desempenho estado-da-arte ($F_1 > 0.9$) para texto português
- Análise de redes revela estruturas coordenadas detectáveis computacionalmente
- Compromisso fundamental: precisão vs. revocação vs. interpretabilidade vs. justiça
- Necessidade de frameworks multimodais para próxima geração de ataques
- Importância de conjuntos de dados eticamente coletados e pesquisa reproduzível

Perguntas e Comentários

Obrigado!

Referências Principais

- Benevenuto, F. & Melo, P. (2024). *Misinformation Campaigns Through WhatsApp and Telegram in Presidential Elections in Brazil*. Communications of the ACM, 67(3), 45-53.
- Kumar, S. et al. (2021). *Detection of COVID-19 Misinformation in Portuguese WhatsApp*. Lecture Notes in Computer Science, 12661, 234-247. Springer.
- DataSenado (2024). *Percepção sobre Fake News e Eleições no Brasil*. Brasília: Senado Federal.
- TSE (2024). *Programa Permanente de Enfrentamento à Desinformação: Relatório Técnico*. Brasília: Tribunal Superior Eleitoral.
- Zhou, X. & Zafarani, R. (2020). *A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities*. ACM Computing Surveys, 53(5), 1-40.