

AVALIAÇÃO DOS MÉTODOS DE MACHINE LEARNING E DE PRÉ PROCESSAMENTO PARA DETECÇÃO DE FAKE NEWS

Stefano X. Soares¹; Roger Monteiro², Daniel Fernando Anderle³; Rodrigo Ramos Nogueira⁴

RESUMO

Com o avanço da Internet a facilidade e a velocidade no compartilhamento de notícias, o problema da disseminação de fake news aflige a sociedade como um todo, afetando cada vez mais o nosso cotidiano. Tendo em vista os problemas causados pela desinformação, este trabalho tem como objetivo o estudo e análise dos métodos de machine learning e de pré-processamento necessários para desenvolver um mecanismo de coleta de dados de forma inteligente a partir de datasets de notícias e a implementação de algoritmos com filtros.

Palavras-chave: Fake news. Notícia falsa. Machine learning. Aprendizado de máquina.

INTRODUÇÃO

No início do século XXI surgiu uma enorme tendência que se segue até os dias atuais: o uso das redes sociais. Cada vez mais pessoas estão conectadas entre si através destas plataformas. Atualmente temos como exemplo de grandes cases de sucesso de redes sociais: o Facebook, o Instagram, o WhatsApp e o Twitter.

Segundo o Statista⁵ o Facebook possuía 608 mil pessoas conectadas no final de 2010. Já no início de 2019, o número de usuários atingiu o montante de 2,375 milhões.

¹ Estudante de Sistema de Informação - Instituto Federal Catarinense - stefano.xavier@hotmail.com

² Estudante de Análise e Desenvolvimento de Sistemas - UNIASSELVI - roger.o.monteiro@gmail.com

³ Professor Instituto Federal Catarinense - Campus Camboriú - daniel.anderle@ifc.edu.br

⁴ Professor Instituto Federal Catarinense - Campus Camboriú - rodrigo.nogueira@ifc.edu.br

⁵ "Number of monthly active facebook users worldwide" - Statista - <https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/>. Acessado em 12 de jun. 2019.

Juntamente com a expansão da Internet nos últimos anos, o número de adeptos das redes sociais também aumentou na mesma proporção, tornando o uso destas ferramentas, cada vez mais, presente do cotidiano das pessoas.

Além de *chats*, grupos de discussões, publicação de fotos, opiniões e sentimentos, um uso que merece destaque nestas redes, por ser muito utilizado, é o de compartilhamento de notícias dos mais diversos assuntos, tais como: esporte, economia, tecnologia, criminalidade, política e outros.

Segundo Nogueira (2018), os sites de notícias são o terceiro maior veículo de informação mais acessado da Internet, perdendo apenas para aplicativos de mensagens e redes sociais.

Juntamente com a importância de textos de notícias e seu compartilhamento das mesmas em redes sociais, vem a ascensão e disseminação das *fake news*.

De acordo com o jornal britânico BBC⁶, durante a disputa presidencial dos EUA entre Donald Jhon Trump e Hillary Diane Rodham Clinton, ocorridas no final de 2016 surgiu uma tendência crescente: a criação, divulgação e compartilhamento de *fake news*.

Durante as eleições o objetivo era claro: enaltecer e promover determinado candidato ou denegrir a imagem do outro, com o intuito de angariar votos a seu favor. Fazendo, assim, surgir o termo “*fake news*” neste período.

Com isso as *fake news* se mostraram potentes influenciadoras do processo eleitoral. Segundo Guarda (2018), no momento de formar sua opinião, o eleitor sofre o impacto de notícias cuja veracidade não é averiguada, criando um juízo em relação aos candidatos e ao processo democrático baseado em notícias falsas. Ainda não é possível afirmar a exata dimensão exercida pelas *fake news* no processo eleitoral, mas é fato que alguma influência elas exercem.

Outros exemplos do impacto causado pela desinformação são os casos apresentados pela BBC⁷ onde onde geraram tanta comoção social que incitaram

⁶ “Como o termo 'fake news' virou arma nos dois lados da batalha política mundial” - BBC - <https://www.bbc.com/portuguese/internacional-42779796> - Acessado em 4 jun. 2019.

⁷ “Três casos de fake news que geraram guerras e conflitos ao redor do mundo” - BBC - <https://www.bbc.com/portuguese/geral-43895609> - Acessado em 6 de jun. 2019

guerra em três situações distintas. A primeira referente a uma falsa denúncia de abusos do exército iraquiano durante o conflito do Kuwait com o Iraque em 1990. A segunda com a divulgação da notícia “Menino crucificado na Ucrânia” inflamando ainda mais o conflito entre tropas ucranianas e forças pró-russas separatistas em 2014. Já a terceira, com a utilização de fotos falsas na crise étnica dos rohingya em Mianmar.

Visando solucionar este problema estão sendo criados centros de discussão pela Mídia onde diversos jornalistas especialistas se reúnem e julgam se uma notícia era verdadeira ou não. Um exemplo é a IFCN⁸ (*Internacional Fact-checking Network*) agência especializada para checagem de notícias que em parceria com o Facebook criou a “Aos Fatos”. Nela, jornalistas identificam as notícias e checam suas fontes para no fim de classificá-las em 7 categorias possíveis: verdadeiro, impreciso, exagerado, distorcido, contraditório, insustentável ou falso.

Segundo Anderle (2017) as aplicações de Smart Cities devem ser voltadas e focadas nos cidadãos e como eles podem usar tecnologias inovadoras, tais como mídias sociais, aplicativos móveis e dados abertos, criando soluções para as questões que interessam a eles e permitindo uma mudança de comportamento.

Tendo em vista o crescente aumento na criação e disseminação de *fake news*, complementado do custo, tempo e impossibilidade de análise na totalidade de tal volume de dados, se tornou clara a necessidade de processamento computacional para identificação destas notícias.

O pré-processamento é muito importante, visto que, os dados sobre sentimento em notícias financeiras serão utilizados para aplicação de técnicas de *Machine Learning*. A etapa de pré-processamento está relacionada ao tratamento do texto, no que diz respeito a eliminação de palavras que não agregam informação ou valor (BENEVENUTO, 2015).

Em processos de aprendizado de máquina, bem como em processo de data warehousing, é considerado que aproximadamente 80% do tempo despendido

⁸ “Internacional fact checking network” - IFCN -

<https://www.poynter.org/tag/international-fact-checking-network/> - Acessado em 6 jun. 2019.

e esforço computacional é utilizado na etapa de pré-processamento dos dados (PYLE, 1999; NAGABUSAHANA,2006; LOSARWAR E JOSHI,2012).

Dito isso, presente trabalho tem como objetivo a avaliação dos métodos de *machine learning* e de pré-processamento necessários para o desenvolvimento de um mecanismo de classificação automática de *fake news*.

Tendo como objetivos específicos: Elaborar um quadro comparativo com as definições de *fake news* apresentando suas classificações, tipologias e principais autores; Levantar as tecnologias de inteligências artificiais já existentes para esta finalidade e os algoritmos utilizados; Buscar e elencar os principais *datasets* existentes de notícias em geral; Buscar e analisar os principais métodos de *machine learning* cabíveis para a detecção de *fake news*, criando um quadro analítico comparativo; Realizar testes de acurácia para determinar qual método e algoritmo se torna mais eficaz; Realização do teste de viabilidade de um detector de *fake news* e respectiva análise.

PROCEDIMENTOS METODOLÓGICOS

Este trabalho será desenvolvido em quatro etapas definidas a partir das metodologias e dos objetivos específicos estabelecidos.

A primeira etapa está relacionada com a pesquisa bibliográfica. Nela será realizada uma revisão sistemática a fim de identificar os principais autores, artigos científicos e material relacionadas ao tema de *fake news* e *machine learning*.

A segunda etapa vincula o trabalho a pesquisa exploratória tendo em vista a sua busca por métodos de machine learning, algoritmos de pré processamento e ferramentas em desenvolvimento existentes na literatura, com a finalidade de detectar *fake news*, bem como realizar análises sobre este tipo de informação.

A terceira etapa está relacionada com a pesquisa aplicada e tecnológica. Esta etapa é responsável pela maior parte do desenvolvimento, tendo em vista que se trata da busca e análise dos métodos de *machine learning* destinados para a análise de linguagem natural existentes nos textos de notícias disponíveis na Web,

para servir de base na criação de quadro analítico comparativo proposto. Em seguida a realização da avaliação das métricas a partir dos datasets encontrados, determinando aqueles mais efetivos na detecção de *fake news*.

Sendo a quarta, e última, etapa ligada a pesquisa qualitativa, voltada para o teste de viabilidade de ferramenta para a detecção de *fake news* e a sua respectiva análise, ou seja, verificar a aceitação dos possíveis usuários da ferramenta.

RESULTADOS ESPERADOS OU PARCIAIS

Este trabalho tem como base o que foi desenvolvido por Roger et al. (2019), que é um sistema de classificação sob o qual os métodos de pré-processamento serão avaliados e aplicados. Além de fundamentar o conceito de *fake news* pelo estado da arte, também será possível, através das técnicas de *machine learning*, desenvolver mecanismos que classifiquem tais notícias de maneira autônoma.

CONSIDERAÇÕES FINAIS

O principal fundamento das *Smart Cities* é oferecer mais bem-estar e qualidade de vida aos cidadãos por meio do uso de tecnologia.

Tendo em vista o avanço da Internet e das redes sociais e por consequência o aumento da facilidade e da velocidade do compartilhamento dessas notícias, os problemas causados pela disseminação de *fake news* tem se agravado cada vez mais, afetando o nosso cotidiano e a sociedade como um todo.

Uma ferramenta que facilitasse a identificação de notícias falsas de forma autônoma seria de extrema importância para garantir um jornalismo sério e real a fim de informar e instruir a população de forma inteligente, evitando os possíveis desdobramentos da desinformação.

REFERÊNCIAS

ANDERLE, Daniel Fernando. **Modelo de conhecimento para representação semântica de smart cities com foco nas pessoas**. <http://btd.egc.ufsc.br/wp-content/uploads/2017/09/Daniel-Fernando-Anderle.pdf>, 2017.

GUARDA, R. F., Ohlson, M. P. and Romanini. **Disinformation, dystopia and post-reality in social media: A semiotic-cognitive perspective**. *Education for Information*, v. 34, n. 3, p. 185–197, 2018.

LOSARWAR, V.; JOSHI, M. **Data Preprocessing in Web Usage Mining**. International Conference on Artificial Intelligence and Embedded Systems. **Anais...** In: ICAIES'2012. Singapore: jul. 2012

MONTEIRO, Roger Oliveira; NOGUEIRA, Rodrigo Ramos; MOSER, Greisse. **Desenvolvimento de um sistema para a classificação de Fakenews com Textos de Notícias em língua Portuguesa**. In: Anais da Escola Regional de Banco de Dados 2019.

NOGUEIRA, Rodrigo Ramos. **O Poder do Data Warehouse em Aplicações de Machine Learning**. São Paulo. Editora NEA. 2018.

PYLE, D. **Data preparation for data mining**. San Francisco, Calif: Morgan Kaufmann Publishers, 1999.

TEIXEIRA, A. **Fake news contra a vida: desinformação ameaça vacinação de combate à febre amarela**. Dissertação. 2019.