

Análise de sites disseminadores de fake news

Davi P. Guimarães, Guilherme M. Moreira, Matheus E. Fagundes

Nilson Mori Lazarin

nilson.lazarin@cefet-rj.br



Introdução

[Lazer et al. 2018].

- Informações fabricadas que imitam notícias confiáveis em formato e estrutura;
- Não possuem o mesmo objetivo nem adotam o mesmo processo editorial;

[Kovic 2018]

- Portais de fake news não garantem a qualidade e confiabilidade da notícia;
- Geralmente publicam notícias com algum tipo de objetivo a ser alcançado;
- Provocam efeitos que vão do inofensivo a extremamente prejudicial;
- Criadas para captar o clique do leitor e fazê-lo compartilhar; Outras podem ter um propósito mais malicioso (e.g. mudar o rumo de uma disputa eleitoral).

Trabalhos relacionados (1/3)

[Tanaka et al. 2010] avalia a credibilidade extraindo informações do conteúdo, do suporte social e do autor da página

- Grau Cobertura, o quanto o conteúdo discorre sobre os tópicos do assunto da página
- Grau de Profundidade do tópico, em que profundidade o assunto da página é tratado pelo texto.

Trabalhos relacionados (2/3)

[Kakol et al. 2017] um modelo para predizer se o conteúdo de uma página web possui credibilidade.

- Através de um estudo empírico com 2041 entrevistados que responderam sobre a credibilidade de 5543 páginas web.
- Uma das maiores dificuldades em se criar modelos para predição de credibilidade é que, quando um humano faz uma avaliação em site, utiliza muitos critérios subjetivos.

Trabalhos relacionados (3/3)

[Dong et al. 2015] busca classificar a veracidade do conteúdo de páginas web, baseado na premissa de que uma fonte que tem poucos fatos falsos é considerada confiável.

- Utilizando um modelo probabilístico foram analisadas 119 mil páginas web e posteriormente parte dos resultados foi verificada manualmente para avaliar a eficiência da técnica.
- Foram analisadas manualmente 100 páginas consideradas confiáveis pelo modelo. Destas, 85 foram consideradas confiáveis pela abordagem manual.

Objetivo

- Identificar características presentes em sites disseminadores de fake news que não estejam presentes em sites confiáveis.
- Contribuir com trabalhos de classificação automática.
- Construção e disponibilização de uma base de URLs de páginas contendo fake news em português brasileiro.

Metodologia

Etapa 1 - Construção da Base de URL

Este trabalho visa analisar fake news em português brasileiro e não foi encontrada nenhuma base de dados pública, o primeiro passo foi construir uma base através do seguinte procedimento:

(1) Eleger sites reconhecidos como checadores de notícias ou desmentidores de boatos. Para tal considerou-se a recomendação realizada por [CGI.br 2018] que indica os sites: E-farsas, Boatos.org e Quatro Cantos;

(2) Através das postagens nos sites checadores, buscou-se armazenar a URL de onde a fake news em questão foi divulgada.

Metodologia

Caso a URL não fosse divulgada no site checador, uma consulta em um buscador era realizada com o texto da fake news.

Nos links retornados pelo buscador, checou-se o conteúdo para observar se tratava-se da mesma notícia avaliada pelo site checador e se a fonte encontrada divulgava o conteúdo como verdadeiro;

Caso ambas observações fossem verdadeiras a URL era adicionada a base de dados.

Metodologia

Etapa 2 - Análise das páginas

Buscou-se comparar as características dos serviços de hospedagem, sendo elas:

- Localização geográfica (obtida através do endereço IP);
- Sistema operacional do servidor;
- Linguagem de programação utilizada e servidor web.

Além disso, analisou-se o conteúdo das páginas em busca do uso excessivo de determinadas palavras, conforme recomendação de [CGI.br 2018] para identificação de boatos.

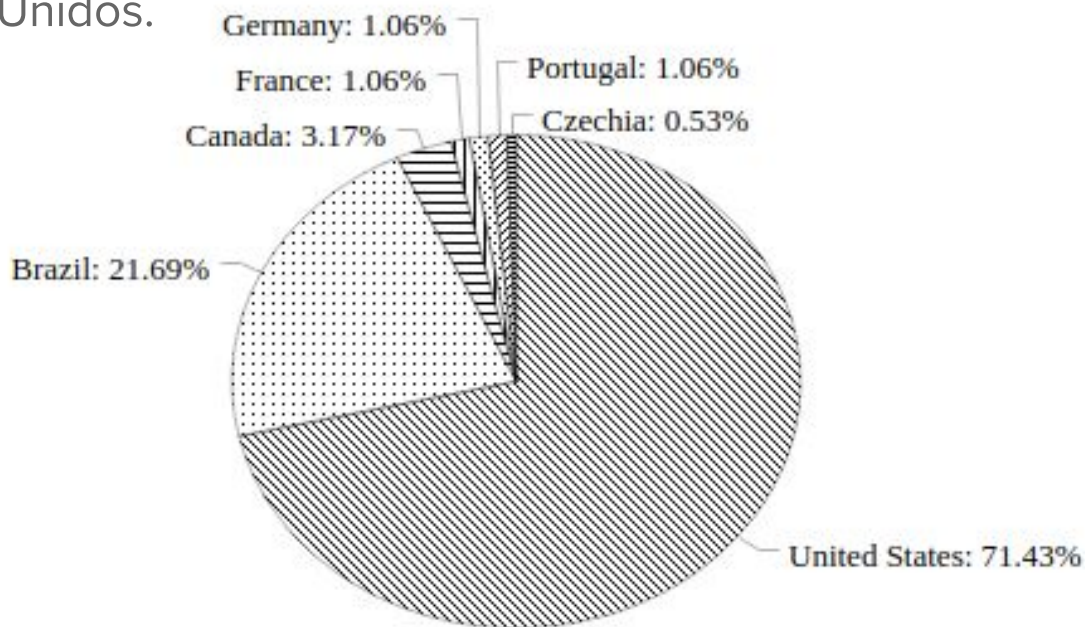
Resultados

A base obtida na fase de busca e catalogação contém 256 URLs de páginas com notícias confirmadas como fake news, através dos sites checadores.

<https://github.com/nilsonmori/urlfakelist>

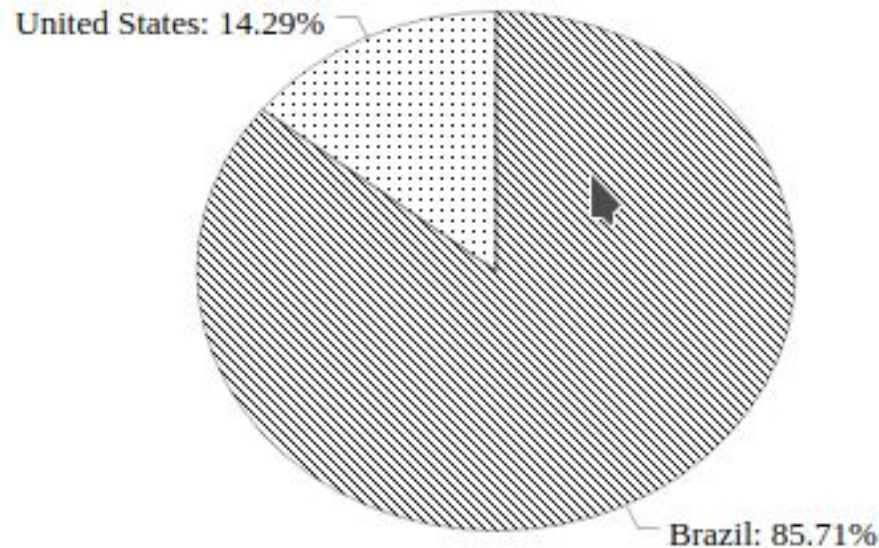
Resultados - País de hospedagem

78.3% dos disseminadores de fake news estão hospedados fora do Brasil, preferencialmente nos Estados Unidos.



Resultados - País de hospedagem

85.7% dos portais de notícias estão hospedados em território nacional.



Resultados - Palavras alarmistas

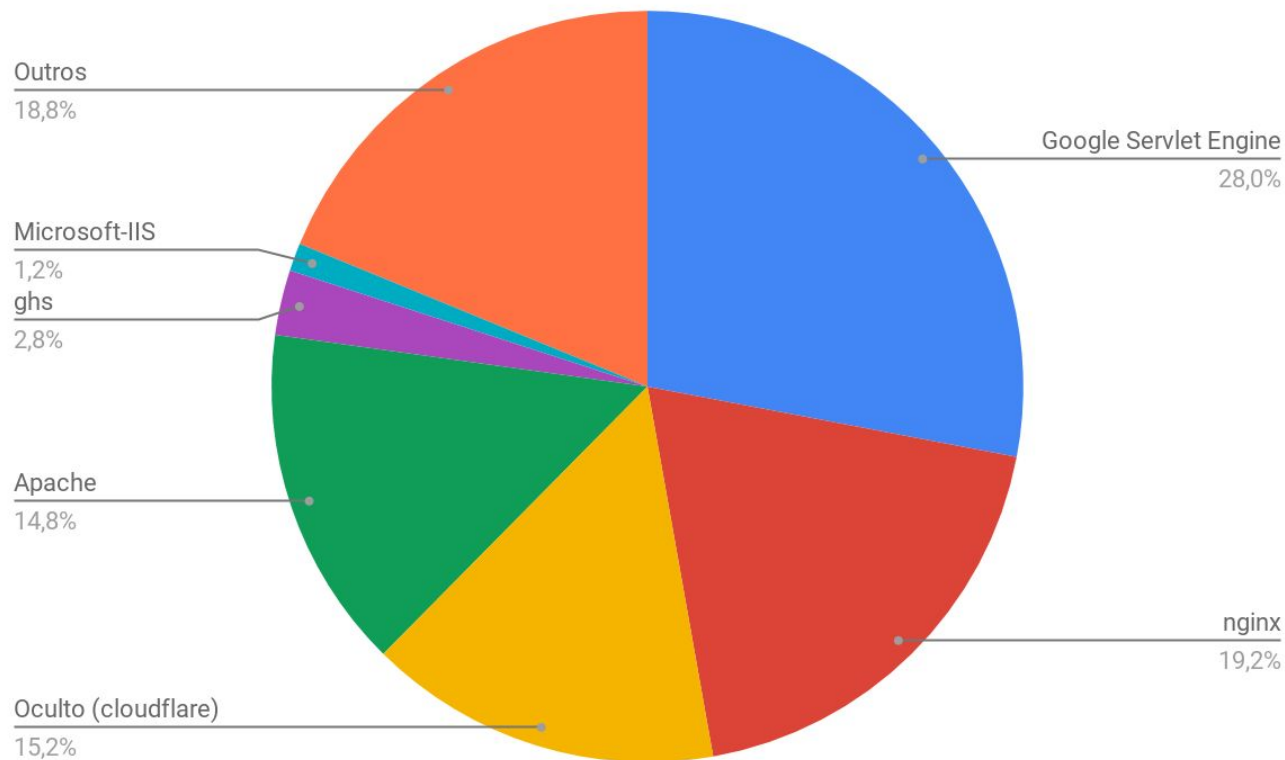
Verificou-se que sites disseminadores de fake news possuem em média 4.5 palavras alarmistas por página, enquanto sites de notícias tendem a não utilizar essas palavras em suas páginas.

As palavras alarmistas utilizadas nesta pesquisa foram

ameaçar, atenção, colabore, compartilhe, contatos, corja, divulgue, enganação, enganado, enganar, espalhem, farsa, grave, gravíssimo, perigo, repassem, sacanagem, urgente e vergonha.

Resultados - Hospedagem

Sites disseminadores de fake news utilizam diversos servidores web, dos quais destacam-se



Conclusão

Algumas características da hospedagem de sites e do conteúdo de páginas:

- Observou-se que em sites disseminadores de fake news a ocorrência de palavras alarmistas é recorrente;
- Na maioria das vezes esses sites estão hospedados fora do Brasil e utilizam diversas tecnologias de hospedagem.

Este trabalho disponibilizou uma base com URLs de páginas contendo notícias em português do Brasil classificadas como fake news por sites checadores.

Trabalhos futuros podem analisar outras características de notícias falsas apontadas por [CGI.br 2018], tais como: Uso excessivo de determinadas tags HTML; Erros ortográficos; Ausência de data da notícia.