

# Detectando Avaliações Falsas na Internet Usando Classificação de Texto

Daniel Atkinson Oliveira<sup>[114054054]</sup>

Universidade Federal do Rio de Janeiro, Rio de Janeiro RJ, Brasil  
danatkho@gmail.com

**Resumo** Nos últimos anos o problema de notícias falsas tem se tornado cada vez mais aparente. A abundância desse tipo de notícia na internet apresenta um grande risco para a maioria dos usuários, especialmente para aqueles que só lêem manchetes. Um outro problema muito relacionado ao de notícias falsas é o de avaliações falsas. Ambos os tipos de publicações têm como objetivo apresentar informações ou opiniões falsas e são difíceis de ser detectados por humanos. No caso das avaliações falsas tanto consumidores quanto prestadores de serviços são afetados. Ambas essas questões, inicialmente muito discutidas no ambiente popular e midiático, estão ganhando cada vez mais atenção no ambiente acadêmico pois esse tipo conteúdo continua a crescer, tornando-o um problema maior, porém ao mesmo tempo, aumentando as possibilidades de estudos. Neste artigo é feita a implementação de um modelo n-grama para detectar automaticamente avaliações falsas. O modelo implementado é baseado no modelo criado por Ahmed et al. [1] e pode ser usado também para detecção automática de notícias falsas. No estudo original duas técnicas de extração de atributos e seis técnicas de classificação por aprendizado de máquina são comparados. Também são testados diferentes valores para os n-gramas e diferentes faixas de atributos. Neste artigo somente será implementado e analisado o caso que resultou na maior taxa de acerto no artigo original.

**Keywords:** avaliações falsas online · notícias falsas · classificação de texto · segurança de redes sociais online.

## 1 Introdução

O experimento que será descrito foi realizado de acordo com a definição do segundo trabalho [4] da disciplina "Recuperação da Informação" [3], da Universidade Federal do Rio de Janeiro. O documento de definição do trabalho [4] diz que o aluno deve "Escolher e estudar um artigo atual da área de Recuperação da Informação. Desenvolver também algum tipo de implementação e experimento relacionado ao artigo selecionado. Escrever um relatório em formato de artigo científico relatando o trabalho desenvolvido e apresentá-lo aos colegas." O artigo de Ahmed et al. [1] foi escolhido, pois, em sua implementação, diversos conceitos que foram abordados em sala de aula são postos em prática, além de ser de ter

apresentado bons resultados para a resolução de um problema muito relevante e atual.

O artigo de Ahmed et al. [1] começa com a apresentação do problema de avaliações falsas online, assim como uma explicação para o surgimento desse fenômeno. Em seguida, é feita uma categorização útil dos tipos de avaliações falsas online: aquelas que têm como propósito apresentar informação falsa sobre o produto ou serviço, seja essa informação positiva ou negativa (tipo 1), aquelas que são puramente direcionadas à marca ou empresa que oferece o serviço ou produto sem apresentar experiência com o produto (tipo 2) e aquelas que não são avaliações ou são textos de marketing, apresentando nenhuma relação ou relação indireta com o produto (tipo 3). É dito no artigo que a categoria mais difícil de identificar é a primeira, e é o foco do artigo, no que diz respeito à detecção de avaliações falsas.

Ahmed et al. [1] apresentam uma categorização de tipos de notícias falsas, argumentando que, pelo fato de, tanto notícias falsas quanto avaliações falsas compartilharem um mesmo atributo, que é o conteúdo falso, seria possível criar um modelo que classificaria tanto notícias falsas quanto avaliações falsas, que é a proposta deles. O modelo proposto por eles é uma combinação de atributos n-gramas de palavras, métricas de frequências de termos e classificação por aprendizado de máquina, sendo testadas duas diferentes técnicas de extração de atributos e seis diferentes técnicas de classificação por aprendizado de máquina. Três diferentes datasets foram usadas para avaliação dos modelos, incluindo notícias falsas e verdadeiras assim como avaliações falsas e verdadeiras. Os resultados obtidos para os datasets usados foram melhores do que as alternativas que já existiam para os mesmos datasets.

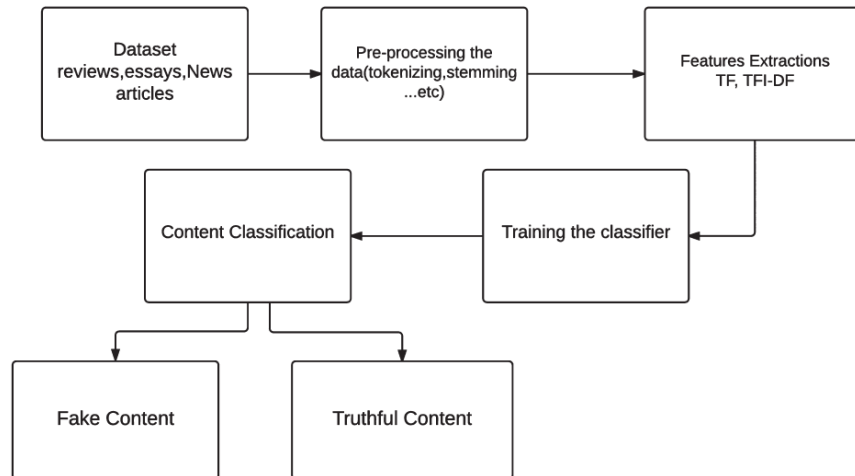
Uma importante distinção a se fazer é a de modelos de detecção baseados em conteúdo versus modelos de detecção baseados em comportamento do avaliador, já que essas são as duas possíveis estratégias a serem tomadas para detecção de avaliações falsas, além da estratégia mista. Um dos principais modelos baseados em conteúdo foi desenvolvido por Ott et al. [2], onde foram usados atributos n-gramas de palavras para detectar avaliações falsas. O dataset, que também foi construído por eles, consiste de 800 avaliações falsas de hotéis, coletadas de Amazon Mechanical Turk, e 800 avaliações verdadeiras coletadas de TripAdvisor. Além disso, metade das avaliações falsas tem cunho positivo e metade tem cunho negativo, o mesmo se aplica às avaliações verdadeiras. Esse dataset veio a ser usado para avaliação de diversos modelos, inclusive, o desenvolvido por Ahmed et al. [1]. A proposta deste artigo é reproduzir o modelo criado por Ahmed et al. [1] que obteve melhor desempenho com o dataset criado por Ott et al. [2].

## 2 Metodologia

### 2.1 Visão Geral

Um fluxograma, representando as etapas percorridas até se chegar na detecção de conteúdo falso, usado por Ahmed et al. [1] pode ser visto na Fig. 1. A par-

tir do dataset de Ott et al. [2] as avaliações passam por uma etapa de pré-processamento, os atributos n-grama de palavras são então adquiridos, uma matriz de atributos por documento é gerada e então o o classificador é treinado.



**Figura 1.** Fluxograma do processo de classificação. Fonte: Ahmed et al. [1].

## 2.2 Pré-processamento

## 3 Resultado

## 4 Discussão

## 5 Conclusão

## Referências

1. Ahmed, H.: Detecting opinion spams and fake news using text classification. Em: Obaidat, M. (ed.) Security and Privacy, Volume 1, Issue 1, e9 (2017) <https://doi.org/10.1002/spy2.9>
2. Ott, M.: Finding deceptive opinion spam by any stretch of the imagination. HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, pp. 309–319. Portland, Oregon (2011)
3. Página da ementa "Recuperação da Informação" <http://dcc.ufrj.br/~giseli/2018-1/ri>  
Rabelo, G.

4. Definição do Trabalho 2  
[http://dcc.ufrj.br/~giseli/2018-1/ri/Definicao\\_Trabalho2.pdf](http://dcc.ufrj.br/~giseli/2018-1/ri/Definicao_Trabalho2.pdf)  
Rabelo, G.