UNIVERSIDADE FEEVALE

FERNANDO AUGUSTO SCHUCH

DETECÇÃO AUTOMÁTICA DE SPAMS DE OPINIÃO EM AVALIAÇÕES DE PRODUTOS

(Título Provisório)

Anteprojeto de Trabalho de Conclusão

Novo Hamburgo
2019

FERNANDO AUGUSTO SCHUCH

DETECÇÃO AUTOMÁTICA DE SPAMS DE OPINIÃO EM AVALIAÇÕES DE PRODUTOS

(Título Provisório)

Anteprojeto de Trabalho de Conclusão de Curso, apresentado como requisito parcial à obtenção do grau de Bacharel em Sistemas de Informação pela Universidade Feevale

Orientador: Prof. Dr. Rodrigo Rafael Villarreal Goulart

Novo Hamburgo
2019

RESUMO

Opiniões sobre bens ou serviços representam uma excelente fonte de informação, tanto para consumidores quanto empresas e fabricantes. Avaliações sobre produtos em *websites* de venda estão sendo cada vez mais consultadas, com o propósito de tomar decisões de compra com base em experiências de outras pessoas. Atualmente, com o aumento no uso das redes sociais, essa prática tem sido adotada por muitos usuários. A confiança nessas avaliações é alta, principalmente em indivíduos entre 18 e 34 anos. Logo, percebe-se que há interesse e necessidade em minerá-las, a fim de acompanhar como está a reputação da marca na Internet. Pelo fato de que *reviews* positivas geralmente significam lucro, enquanto negativas afetam a reputação dos produtos, este cenário motiva a postagem de opiniões falsas, buscando persuadir o consumidor a tomar decisões erradas. Essa atividade, conhecida como spam de opinião, é uma vertente da mineração de opiniões que recebeu atenção somente a partir de 2008. Apesar de já existirem estudos nessa área, ela tem sido pouco abordada na língua portuguesa. Portanto, há escassez de exemplos anotados, ou seja, classificados como spam ou não-spam para a criação de algoritmos de detecção. Esta pesquisa realizará experimentos em um corpus de avaliações sobre mercadorias, objetivando identificar opiniões de usuários que não possuem o produto sobre o qual estão opinando. Para tanto, serão propostas técnicas de aprendizado de máquina para anotação automática das mesmas. Com isso, será possível analisar como este tipo de spam impacta na reputação online dos produtos.

Palavras-chave: Mineração de opiniões. Spams de opinião. Anotação de corpus. *Reviews*. Aprendizado de Máquina.

SUMÁRIO

MOTIVAÇÃO 5

OBJETIVOS 8

METODOLOGIA 9

CRONOGRAMA 11

BIBLIOGRAFIA 12

MOTIVAÇÃO

As redes sociais, devido à evolução da Internet e do crescimento no uso de *smartphones*, impulsionaram a geração de opiniões sobre produtos, serviços, estabelecimentos, pessoas e fatos. Atualmente, há uma grande facilidade em publicar e consultar tais informações nesse meio (CARDOSO, 2017). Liu (2012) afirma que as opiniões de mídias sociais têm sido utilizadas para diversos fins, desde decisões de compra, marketing e design de produtos, até escolhas em eleições políticas. Há, portanto, um crescente interesse em minerar opiniões na Web, já que o conteúdo gerado por usuários contém informações valiosas, as quais podem ser exploradas em diferentes aplicações.

Cada vez mais consumidores optam por tomar decisões de compra de produtos baseadas em avaliações e experiências postadas por outras pessoas em *websites*, o que motiva comerciantes a criar *reviews* falsas, tanto para melhorar sua reputação quanto para atacar competidores (LIN et al., 2014). Mukherjee, Liu e Glance (2012) complementam que se alguém quer comprar um produto, irá ler suas avaliações. Se a maioria for positiva, tenderá a comprá-lo, mas se grande parte for negativa, provavelmente o consumidor escolherá outra mercadoria. Esse cenário proporciona fortes incentivos para a atividade conhecida como spam de opinião.

Estudos da Bright Local (2018) sobre o uso de estabelecimentos por consumidores apontam que, no Reino Unido, 91% das pessoas entre 18 e 34 anos de idade confiam em *reviews* online tanto quanto recomendações pessoais. Enquanto que 40% dos entrevistados dizem que não utilizariam um serviço caso leiam uma opinião negativa. Além disso, 89% dos entrevistados dessa faixa etária afirmam já terem lido avaliações enganosas. Devido ao elevado número de falsificações, empresas como a Amazon já se preocupam em criar esforços para reduzir o número de spams e *spammers*, ou seja, usuários que escrevem opiniões falsas. Em 2015, a companhia entrou com ações judiciais contra sites que vendiam *reviews* fraudulentas de consumidores para comerciantes da Amazon (BISHOP, 2015).

De acordo com Jindal e Liu (2008), spams de opiniões podem ser divididos em três tipos:

- **Falsos:** *Reviews* positivas para produtos ou serviços que não as merecem, bem como opiniões negativas que servem para difamar a reputação do alvo;

- **Sobre marcas:** Avaliações que não falam sobre as características do produto em si, mas somente sobre a marca, fabricante ou vendedor;
- **Não-avaliações:** propagandas e outros textos irrelevantes, como perguntas ou respostas.

Estes tipos de spam diferem dos encontrados em e-mails ou em páginas Web. O primeiro refere-se a links para outros sites indesejados, enquanto que o segundo é caracterizado pela inserção de palavras populares em sites, para que motores de pesquisa os marquem como relevantes. Ambos têm sido amplamente estudados (LIU, 2012).

Jindal e Liu (2008) foram os primeiros autores a escrever sobre spam de opinião, o que mostra a recenticidade dessa área. Desde então, como afirma Sandulesco e Ester (2015), diversos trabalhos têm sido publicados com o objetivo de identificar atividades relacionadas ao spam de opinião. Segundo eles, a maioria dessas pesquisas seguem duas direções: análises de características comportamentais ou de textos. Do ponto de vista comportamental, podem ser utilizadas informações como a data da *review*, seu *rating* médio (expresso em número de pontos ou estrelas), de onde a mesma foi postada e assim por diante. Já do lado textual, sua análise se baseia em extrair pistas do conteúdo da avaliação, desde padrões de texto até a frequência das palavras. Apesar dessa distinção, a maioria dos estudos realizados utilizaram-se de ambas as vertentes para maximizar a eficácia dos resultados, como Jindal e Liu (2008), Yuan et al. (2016), Lim et al. (2010), Mukherjee, Liu e Glance (2012) e Cardoso (2017). Embora existam pesquisas que focaram em características comportamentais (LI, 2015; MUKHERJEE et al., 2013) e linguísticas (SANDULESCO; ESTER, 2015; OTT et al., 2011) com mais ênfase.

Os estudos citados acima contribuíram para a detecção automáticas de spams na língua inglesa, mas poucos esforços foram empreendidos para a língua portuguesa. Até o momento, tem-se apenas o conhecimento de que Costa, Benevenuto e Merschmann (2013) analisaram um corpus de spam em português para este fim. Os autores desenvolveram um modelo de classificação automática de três classes de spam sobre estabelecimentos comerciais de uma rede social: poluidoras, bocas-sujas e propagandas. Utilizaram de técnicas das duas direções descritas por Sandulesco e Ester (2015), obtendo consideráveis 84% de acerto na identificação de opiniões spam.

A presente pesquisa visa complementar os trabalhos existentes, através da análise de opiniões em português. Pretende-se identificar spams que indiquem o desejo dos usuários comprarem um determinado produto, mas que ainda não adquiriram o mesmo. A Figura 1

elucida uma *review* com essa característica. Apesar de afirmar que o produto é “muito bom e muito bonito”, há indícios que o usuário não o possui. Logo, esta avaliação não é útil para outros consumidores que se interessarem pela mesma mercadoria.

Figura 1 – Avaliação de usuário que não possui a mercadoria

Fonte: Buscapé (2013)

Assim como as opiniões analisadas por Costa, Benevenuto e Merschmann (2013), essa classe de spam relacionada com o tipo não-avaliação, definido por Jindal e Liu (2008), ganhará a nomenclatura de opiniões de *não-possuidores*. *Reviews* como essas alteram o *rating* médio dos produtos, afetando também o comportamento do comprador, o qual baseia-se em avaliações que não demonstram a real qualidade da mercadoria. Após revisão bibliográfica, não foram encontrados corpora anotados como possuidores e não-possuidores. Portanto, a anotação de um corpus com a finalidade de criar modelos de Aprendizado de Máquina para detecção automática, bem como analisar as implicações desse tipo de spam na língua portuguesa, é algo ainda não explorado.

OBJETIVOS

Objetivo geral

O objetivo desta pesquisa é identificar opiniões de usuários de um corpus com avaliações de produtos que se caracterizam como não-possuidores, a fim de elaborar um modelo de detecção automática das mesmas, analisando o seu impacto na reputação dos produtos.

Objetivos específicos

- Estabelecer uma metodologia para anotação manual e parcial do corpus de opiniões em spams de não-possuidores;
- Identificar uma métrica para avaliação do grau de confiabilidade da anotação;
- Investigar técnicas de Aprendizado de Máquina para anotação automática do restante do corpus;
- Analisar o impacto desse tipo de opinião no *rating* médio dos produtos.

METODOLOGIA

De acordo com Prodanov e Freitas (2013), o presente estudo é caracterizado como pesquisa aplicada, já que propõe a geração de conhecimentos a serem aplicados de forma prática na solução de problemas específicos. Além disso, possui caráter exploratório, dado que será realizado um levantamento bibliográfico a fim de levantar informações e exemplos que estimulem a compreensão sobre o problema proposto. Do ponto de vista dos procedimentos técnicos, o trabalho é classificado como pesquisa-ação, pelo "interesse coletivo na resolução de um problema ou suprimento de uma necessidade." (PRODANOV; FREITAS, 2013).

Para o desenvolvimento desta pesquisa, buscou-se por corpora sobre opiniões de bens ou serviços disponibilizados por outros autores. Em Hartmann et al. (2014), foi construído um corpus com 85.910 *reviews* extraídas do Buscapé, um site de serviços em português, no qual é possível postar vantagens e desvantagens sobre produtos, serviços e empresas (HARTMANN et al., 2014). As avaliações foram extraídas com o intuito de normalizá-las lexicamente, ou seja, corrigindo erros de escrita, abreviações e gírias comuns na Internet. Este corpus foi escolhido para análise de spams de opinião pelos seguintes motivos: 1) ser completamente em português; 2) grande número de *reviews*; 3) acesso na íntegra, aos textos das opiniões e metadados sobre a avaliação, como ID do usuário, *rating* e data de postagem.

"Em geral, detecção de spams de opinião pode ser formulada como um problema de classificação entre duas classes, falsas e não-falsas. O aprendizado supervisionado é naturalmente aplicável." (LIU, 2012, p. 115, tradução nossa). Porém, devido à enorme quantidade de opiniões, será utilizado o aprendizado semi-supervisionado. Zhu, Ghahramani e Lafferty (2003) argumentam que exemplos rotulados, em Aprendizado de Máquina, são custosos de conseguir, portanto a combinação entre dados rotulados e não rotulados tem importância central nessa área.

Primeiramente, será proposto uma metodologia para classificar uma parcela dos *reviews* do Buscapé em duas classes, opiniões de possuidores ou não-possuidores. Para isso, serão selecionados voluntários, a fim de maximizar o número de anotações manuais. Cada *review* será classificada por mais de uma pessoa, para que suas conclusões possam ser comparadas entre si. Através de métricas para cálculo de concordância, poderemos eliminar as opiniões que geraram discordância entre os voluntários, aumentando assim, a confiança nos exemplos anotados.

Na próxima etapa, serão pesquisadas técnicas de Aprendizado de Máquina, as quais poderão ser aplicadas para encontrar características nos exemplos previamente anotados, como propósito de gerar um modelo para classificação automática do restante das *reviews*. Ao final, será analisado o nível de assertividade do modelo proposto, além de avaliar como as opiniões de não-possuidores podem interferir no *rating* médio dos produtos do Buscapé. Número esse, responsável por sinalizar a reputação das mercadorias no *website*.

CRONOGRAMA

Trabalho de Conclusão I

Etapa	Meses			
	Mar	Abr	Maio	Jun
Escrita anteprojeto				
Revisão anteprojeto				
Revisão da literatura: spam de opiniões				
Revisão da literatura: metodologia para anotação do corpus				
Aplicação da metodologia para anotação do corpus				
Escrita TCC I				
Revisão TCC I				
Entrega TCC I				

Trabalho de Conclusão II

Etapa	Meses			
	Ago	Set	Out	Nov
Avaliação de concordância entre as anotações				
Identificação e aplicação de técnicas de Aprendizado de Máquina				
Análise do impacto das opiniões spam na reputação dos produtos				
Escrita TCC II				
Revisão TCC II				
Entrega TCC II				

BIBLIOGRAFIA

BISHOP, Todd. Amazon files first-ever suit over fake product reviews, alleging sites sold fraudulent praise. **Geek Wire**, Seattle, abr. 2015. Disponível: <https://www.geekwire.com/2015/amazon-files-first-ever-suit-over-fake-reviews-alleging-calif-man-sold-fraudulent-praise-for-products/>. Acesso em: 17 mar. 2019.

BUSCAPÉ. **Puma Borussia Dortmund Jogo 2012/13 III Manga Longa Masculino**. 2013. Disponível em: <https://www.buscape.com.br/avaliacoes/puma-borussia-dortmund-jogo-2012-13-iii-manga-longa-masculino/negativas>. Acesso em: 5 abr. 2019.

CARDOSO, Emerson Freitas. **Filtragem Automática de Opiniões Falsas: Comparação Compreensiva dos Métodos baseados em Conteúdo**. 2017. 88 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de São Carlos (UFSCar), Sorocaba, SP, 2017. Disponível em: <https://repositorio.ufscar.br/handle/ufscar/9141?show=full>. Acesso em: 17 mar. 2019.

COSTA, Helen; BENEVENUTO, Fabrício; MERSCHMANN, Luiz H. C. Detecting Tip Spam in Location-based Social Networks. **Proceedings of the ACM Symposium on Applied Computing**, Coimbra, mar. 2013. Disponível em: <https://homepages.dcc.ufmg.br/~fabricio/download/sac2013.pdf>. Acesso em: 23 mar. 2019.

HARTMANN, Nathan S. et al. A Large Corpus of Product Reviews in Portuguese: Tackling Out-Of-Vocabulary Words. **Language Resources and Evaluation Conference**, Reykjavik, mai. 2014. Disponível em: <https://bdpi.usp.br/bitstream/handle/BDPI/45557/2483927.pdf>. Acesso em: 17 mar. 2019.

JINDAL, Nitin; LIU, Bing. Opinion Spam and Analysis. **Proceedings of the 2008 international conference on web search and data mining**, Palo Alto, fev. 2008. Disponível em: < https://www.cs.uic.edu/~liub/FBS/opinion-spam-WSDM-08.pdf>. Acesso em: 17 mar. 2019.

LI, Huayi et al. Analyzing and Detecting Opinion Spam on a Large-Scale Dataset via Temporal and Spatial Patterns. **Proceedings of the Ninth International AAAI Conference on Web and Social Media**, Oxford, mai. 2015. Disponível em: <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/download/10534/10461>. Acesso em: 16 mar. 2019.

LIM, Ee-Peng et al. Detecting Product Review Spammers using Rating Behaviors. **International Conference on Information and Knowledge**, Toronto, out. 2010. Disponível em: <https://www.cs.uic.edu/~liub/publications/cikm-2010-final-spam.pdf>. Acesso em: 25 mar. 2019.

LIN, Yuming et al. Towards Online Anti-opinion Spam: Spotting Fake Reviews from the Review Sequence. **Internacional Conference on Advances in Social Networks Analysis and Mining**, Pequim, ago. 2014. Disponível em: <https://ieeexplore.ieee.org/document/6921594>. Acesso em: 17 fev. 2019.

LIU, Bing. **Sentiment Analysis and Opinion Mining**. Toronto: Morgan & Claypool. 2012. 165 p.

LOCAL Consumer Review Survey. **Bright Local**, Reino Unido, 2018. Disponível em: <https://www.brightlocal.com/learn/local-consumer-review-survey/#local-business-review-habits>. Acesso em: 17 mar. 2019.

MUKHERJEE, Arjun et al. Spotting Opinion Spammers using Behavioral Footprints. **Conference on Knowledge Discovery and Data Mining**, Chicago, ago. 2013. Disponível em: <https://www.cs.uic.edu/~liub/publications/KDD-2013-Arjun-spam.pdf>. Acesso em: 23 mar. 2019.

MUKHERJEE, Arjun; LIU, Bing; GLANCE, Natalie. Spotting Fake Review Groups in Consumer Reviews. **World Wide Web Conference**, Lyon, abr. 2012. Disponível em: <https://www.cs.uic.edu/~liub/publications/WWW-2012-group-spam-camera-final.pdf>. Acesso em: 17 mar. 2019.

OTT, Myle et al. Finding Deceptive Opinion Spam by Any Strecth of the Imagination. **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics**, Portland, jun. 2011. Disponível em: <https://www.aclweb.org/anthology/P11-1032>. Acesso em: 17 mar. 2019.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar de. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico**. 2. ed. Novo Hamburgo, RS: Feevale, 2013.

SANDULESCO, Vlad; ESTER, Martin. Detecting Singleton Review Spammers Using Semantic Similarity. **World Wide Web Conference**, Florença, mai. 2015. Disponível em: <http://www2015.wwwconference.org/documents/proceedings/companion/p971.pdf>. Acesso em: 25 mar. 2019.

YUAN, Yuan et al. Interpretable and Effective Opinion Spam Detection via Temporal Patterns Mining across Websites. **IEEE International Conference on Big Data**, Washington, dez. 2016. Disponível em: <http://dulcimer.cse.lehigh.edu/~sxie/paper/bigdata16a.pdf>. Acesso em: 17 mar. 2019.

ZHU, Xiaojin; GHAHRAMANI, Zoubin; LAFFERTY, John. Semi-supervised Learning Using Gaussian Fields and Harmonic Functions. **Proceedings of the Twentieth International Conference on Machine Learning**, Washington, 2003. Disponível em: <https://www.aaai.org/Papers/ICML/2003/ICML03-118.pdf>. Acesso em: 31 mar. 2019.