UNIVERSIDADE
LUSÓFONA

Deteção de reviews falsas em plataformas de e-commerce

Trabalho Final de Curso

Relatório Intercalar 1º Semestre

Aluno: Nuno Santinhos
Orientadora: Sofia Fernandes

Trabalho Final de Curso | LCiD | Dezembro 2024

www.ulusofona.pt

Direitos de cópia

Este documento foi gerado com o processador (pdf/Xe/Lua)LaTeX e o modelo ULThesis (v1.0.0) [1].

Resumo

O crescimento do *e-commerce* trouxe um aumento significativo nas avaliações, que desempenham um papel determinante nas decisões de compra dos consumidores. Contudo, o fenómeno das *fake reviews* compromete a confiança nas plataformas digitais, prejudicando tanto os consumidores como as empresas. Este trabalho propõe a aplicação de técnicas de aprendizagem automatizada, como o algoritmo *K-Means*, e de modelos avançados baseados em redes neuronais, com o objetivo de identificar padrões associados a avaliações falsas. Recorrendo a um conjunto de dados da *Amazon*, recolhido entre 1996 e 2023, serão analisadas métricas comportamentais e linguísticas que possibilitam a distinção entre avaliações genuínas e falsas. Este estudo sublinha a relevância de métodos inovadores e integrados para enfrentar os desafios da manipulação de avaliações, promovendo um ambiente de comércio digital mais mais transparente e fiável.

Palavras-chave: Avaliações falsas, comércio eletrónico, aprendizagem automatizada, *K-Means*, análise comportamental, características linguísticas, deteção de avaliações falsas, redes neuronais

Abstract

The growth of e-commerce has led to a significant increase in reviews, which play a decisive role in consumers' purchasing decisions. However, the phenomenon of fake reviews undermines trust in digital platforms, harming both consumers and businesses. This study proposes the application of machine learning techniques, such as the K-Means algorithm, and advanced neural network-based models to identify patterns associated with fake reviews. Using an Amazon dataset collected between 1996 and 2023, behavioral and linguistic metrics will be analyzed to distinguish genuine reviews from fake ones. This research highlights the importance of innovative and integrated methods to address the challenges of review manipulation, promoting a more transparent and reliable digital commerce environment.

Keywords: Fake reviews, e-commerce, machine learning, K-Means, behavioral analysis, linguistic features, fake review detection, neural networks

Índice

Lista de Figuras

Lista de Tabelas

Introdução

1.1 Contexto e Relevância

As *fake reviews*, também conhecidas como avaliações falsas, são comentários criados com o objetivo de manipular a perceção de consumidores em plataformas de e-commerce. Esses comentários, que podem ser positivos ou negativos, são um fator determinante na decisão de uma compra, tornando-se um elemento crucial nos mercados digitais. As avaliações positivas podem atrair novos clientes e gerar lucros, enquanto as avaliações negativas podem afastar consumidores e prejudicar a reputação das marcas.

As *fake reviews* não são apenas feitas por humanos, *bots* automatizados também têm sido utilizados para criar avaliações falsas. Esses bots conseguem gerar grandes volumes de comentários imitando o estilo de escrita humana. Estudos indicam que cerca de 30% das avaliações no *TripAdvisor* são falsas, frequentemente criadas por indivíduos sem qualquer experiência real com o produto ou serviço [2]. Este tipo de manipulação prejudica tanto os consumidores quanto a integridade dos negócios.

Além disso, classificar uma avaliação como falsa ou verdadeira é uma tarefa complexa. Muitos comentários falsos apresentam uma aparência genuína, de modo a sua identificação. Os métodos manuais de deteção de *fake reviews* alcançam uma precisão limitada, tornando necessário o recurso a abordagens mais sofisticadas [3].

1.2 Impactos e Desafios

Os impactos das *fake reviews* vão além de prejuízos financeiros. Elas comprometem a confiança dos consumidores nas plataformas, conduzindo a decisões equivocadas. Um cliente pode evitar a aquisição de um produto ou serviço com base em experiências negativas falsas, resultando em perdas para empresas. Por outro lado, comentários positivos enganosos podem levar consumidores a adquirir produtos que não correspondem às expectativas, resultando em frustração e insatisfação.

Para combater este problema, vários desafios vão ter de ser superados. Um dos principais é a dificuldade em identificar quais avaliações são realmente falsas (*ground truth*). A criação de bases de dados confiáveis e bem anotadas é um processo caro e demorado, frequentemente sujeito a erros humanos. A diversidade linguística nas avaliações agrava ainda mais o problema, dado que muitos métodos se focam em apenas uma língua, limitando a eficácia em contextos multilíngues. Além disso, limitações tecnológicas e computacionais restringem o processamento de grandes volumes de dados, dificultando a exploração completa do problema.

2 - Pertinência e Viabilidade

2.1 Impacto das *Fake Reviews* no Comércio Digital

As avaliações online tornaram-se um componente essencial no comércio digital, desempenhando um papel importante nas decisões dos consumidores. Contudo, a manipulação crescente destas avaliações por meio de *fake reviews* compromete a confiança nas plataformas digitais, pois prejudica a transparência e distorce a concorrência [2]. Este problema afeta tanto grandes empresas como pequenas, especialmente aquelas que dependem fortemente da credibilidade das opiniões para atrair clientes.

Garantir que as avaliações disponíveis sejam genuínas é fundamental para preservar um ambiente de confiança. A exposição dos consumidores a opiniões manipuladas pode levar a decisões de compra inadequadas. A evolução constante dos padrões de avaliações falsas torna o problema ainda mais complexo, exigindo soluções inovadoras e adaptáveis [3].

2.2 Relevância e Benefícios da Solução

O combate às *fake reviews* não se limita à sua deteção, estando também relacionado com a promoção de práticas comerciais justas e a proteção dos consumidores contra informações enganosas. Este tema é particularmente relevante num cenário onde o impacto destas avaliações vai além do comércio digital, influenciando o turismo, os serviços e as plataformas de recomendação. Estudos mostram que práticas enganosas podem prejudicar severamente tanto consumidores como empresas, reforçando a necessidade de métodos eficazes para enfrentar este problema [2].

Além disso, a resolução deste problema contribui para o avanço de outras tecnologias, como sistemas de recomendação mais confiáveis e análises preditivas baseadas em dados limpos e fiáveis. Essas melhorias não só aumentam a confiança dos consumidores nas plataformas, mas também incentivam práticas comerciais éticas, promovendo um mercado mais sustentável e competitivo [3].

2.3 Inovação Proposta neste Trabalho

Este trabalho irá se destingir de estudos, anteriormente, feitos pois irá ser explorado um *dataset* mais recente e, consequentemente, irá ter padrões atuais de avaliações falsas. Adicionalmente, vai se destacar o uso de novas combinações de diferentes tipos de informação em modelos, que é uma estratégia que ainda não foi, fortemente, explorada.

3 - Benchmarking

3.1 Extração de *Features*

A eficácia dos modelos depende também da qualidade e relevância das *features* extraídas. Estas que são atributos ou métricas extraídas dos dados que servem como entrada para modelos de *machine learning*.

Maximum Number of Reviews, este cálculo é feito contabilizando todas as avaliações publicadas e identificando o maior número em qualquer intervalo temporal, como um dia, uma semana ou um mês. O cálculo é feito com base na fórmula 1 [4].

$$\frac{MaxRev(a)}{\max(MaxRev)} \tag{1}$$

Onde *a* é o utilizador que está a ser analisado, o *MaxRev(a)* representa o número máximo de avaliações pelo utilizador num intervalo de tempo, por exemplo num dia, o *MaxRev* é o conjunto de número máximo de avaliações de todos os utilizadores e *max(MaxRev)* é o valor máximo desse conjunto.

Percentage of Positive Reviews avalia a percentagem de avaliações positivas realizadas por um utilizador. Este valor é calculado dividindo o número de avaliações classificadas como positivas pelo número total de avaliações publicadas. Utilizadores que frequentemente publicam avaliações positivas podem ser considerados suspeitos, uma vez que tal comportamento pode indicar uma possível manipulação de avaliações O cálculo é feito com base na fórmula 2[4].

$$\frac{\sum_{x=1}^{|R_a|} |\star(r_x) \in \{4,5\}|}{|R_a|} \tag{2}$$

Na fórmula apresentada, R_a representa o conjunto de avaliações realizadas pelo utilizador a, enquanto $|R_a|$ corresponde ao número total de avaliações presentes nesse conjunto. A variável $\star(r_x)$ refere-se à pontuação atribuída a uma avaliação específica r_x, e $\{4,5\}$ identifica os valores que são considerados positivos para efeitos do cálculo.

Average Review Length calcula o comprimento médio das avaliações realizadas por um utilizador. Este cálculo é realizado somando o número total de palavras em todas as avaliações e dividindo pelo número total de avaliações publicadas. Estudos demonstram que avaliações falsas tendem a ser mais curtas e menos detalhadas do que as genuínas, o que torna esta *feature* particularmente útil para as distinguir O cálculo é feito com base na fórmula 3 [4].

$$\begin{cases} 1, & \text{se } len(r_a) < X \\ 0, & \text{caso contrário.} \end{cases} \tag{3}$$

Na fórmula, $len(r_a)$ representa o comprimento, em número de palavras, de uma única avaliação r_a. O parâmetro X é um valor configurável que estabelece o limiar entre avaliações curtas e longas. Caso o comprimento de r_a seja inferior a X, a avaliação é classificada como curta (1); caso contrário, é classificada como longa (0).

Burstness mede a frequência com que as avaliações são publicadas em períodos curtos com uma intensidade anómala. Para calcular esta *feature*, analisa-se a distribuição temporal das publicações de um utilizador. Padrões de explosões de atividade, como a publicação de um número elevado de avaliações num curto intervalo de tempo, estão frequentemente associados a comportamentos suspeitos, como a realização de avaliações falsas. Esta *feature* é particularmente eficaz para identificar redes coordenadas de *bots* ou ações organizadas em grupo O cálculo é feito com base na fórmula 4 [4].

$$\begin{cases} 1, & \text{se } \sum_{x=1}^{|R_a|} |r_x \in R_a \cap (\text{últimas 24 horas})| > X \\ 0, & \text{caso contrário.} \end{cases} \tag{4}$$

Na fórmula apresentada, $r_x \in R_a$ refere-se a uma avaliação específica dentro do conjunto R_a, que representa todas as avaliações realizadas por um utilizador a. O termo $R_a \cap$(últimas 24 horas) indica a interseção entre o conjunto R_a e as avaliações publicadas nas últimas 24 horas. A soma $\sum_{x=1}^{|R_a|}$ calcula o número total de avaliações que cumprem este critério temporal. Finalmente, o parâmetro X é um limiar configurável que determina se a frequência de avaliações no período especificado excede o limite pré-definido.

Estas *features*, juntamente, com características linguísticas e do produto, fornecem aos modelos uma grande base para identificar padrões que indicam avaliações falsas, isto quando temos o *ground truth*. No entanto, o desenvolvimento de modelos precisos também exige métricas adequadas para avaliar o seu desempenho. A precisão (*accuracy*), a pontuação F1 (F1 Score), são amplamente utilizadas para medir a capacidade do modelo de diferenciar entre avaliações falsas e genuínas, enquanto a perda logarítmica (*log-loss*) é usada para avaliar a confiança nas precisões probabilísticas [2].

3.2 Modelos de Machine Learning

Os métodos de *machine learning* para deteção de *fake reviews* dividem-se, de forma geral, em duas categorias principais: supervisionados e não supervisionados.

3.2.1 Modelos Supervisionados

Os métodos supervisionados utilizam um conjunto de dados rotulados para treinar o modelo, onde cada *review* é previamente classificado como falsa ou genuína. Este tipo de abordagem requer uma base de dados de treino rotulados (*ground truth*), o que pode ser um desafio, já que a criação deste tipo de dados é cara e demorada [2]. Exemplos de modelos supervisionados incluem *Support Vector Machines* (SVM), Regressão Logística e *Random Forest*. Estes modelos funcionam bem quando têm acesso a um grande volume de dados rotulados, mas tendem a ser menos eficazes em situações onde os dados estão desbalanceados ou são escassos.

3.2.2 Modelos Não Supervisionados

Por outro lado, os métodos não supervisionados não necessitam de dados rotulados. Em vez disso, baseiam-se na análise de padrões ou agrupamentos nos dados. Estes métodos são úteis em situações onde não é possível obter etiquetas confiáveis, como em plataformas que geram milhares de avaliações diariamente.

Uma das técnicas mais utilizadas neste contexto é o *K-Means*, que organiza os dados em grupos (*clusters*) com base na similaridade das suas características. No caso da deteção de *fake reviews*, o *K-Means* pode ser usado para identificar agrupamentos de avaliações que partilhem padrões suspeitos, como textos semelhantes ou comportamentos anómalos por parte de utilizadores [3]. Apesar de serem menos dependentes de dados anotados, os métodos não supervisionados enfrentam desafios na interpretação dos agrupamentos e podem produzir resultados menos precisos quando aplicados isoladamente.

3.2.3 Modelos Avançados

Modelos mais complexos têm sido explorados para análise do textos das *reviews*, nomeadamente redes neuronais convolucionais (CNN) e recorrentes (RNN e LSTM) [5].As CNN identificam padrões no texto, como repetições ou frases genéricas. As RNN e LSTM destacam-se na análise de relações entre palavras em textos mais longos. Embora estas redes pertençam à categoria de modelos supervisionados, algumas implementações têm combinado técnicas não supervisionadas para melhorar a robustez da deteção [2].

BERT e *RoBERTa* são modelos de linguagem baseados em *Transformes*, conhecidos pela capacidade de capturar o contexto bidirecional das palavras, o que os torna eficazes em tarefas complexas de classificação [6]. Um estudo destacou que o *RoBERTa* alcançou uma precisão de 91,2% na deteção de *reviews* falsas, superando os diversos métodos [3].

3.3 Integração de Métodos

Embora os métodos supervisionados e não supervisionados apresentem diferenças claras, eles não são mutuamente exclusivos. Muitas abordagens atuais combinam técnicas de ambos os tipos para maximizar a precisão e minimizar as limitações individuais. Esta integração tem sido essencial para lidar com os desafios do volume crescente de dados e a constante evolução das estratégias por parte dos *spammers*. Além disso, a flexibilidade desta combinações permite que os modelos sejam ajustados para diferentes contextos e línguas [3].

4 - Background Teórico

4.1 *K-Means* na Identificação de Padrões

O algoritmo *K-Means* é uma técnica de agrupamento que organiza os dados em grupos (*clusters*) através do cálculo de centróides que representam os centros desses grupos. Cada observação é atribuída ao *cluster* cujo centróide está mais próximo, e o algoritmo ajusta iterativamente os centróides até que a formação dos *clusters* se estabilize. No contexto das avaliações falsas, esta técnica é eficiente para identificar padrões comuns em grupos de avaliações, permitindo destacar semelhanças nos comportamentos dos utilizadores. Na figura 1, gerada com dados *random*, podemos observar o funcionamento do algoritmo *K-Means*.

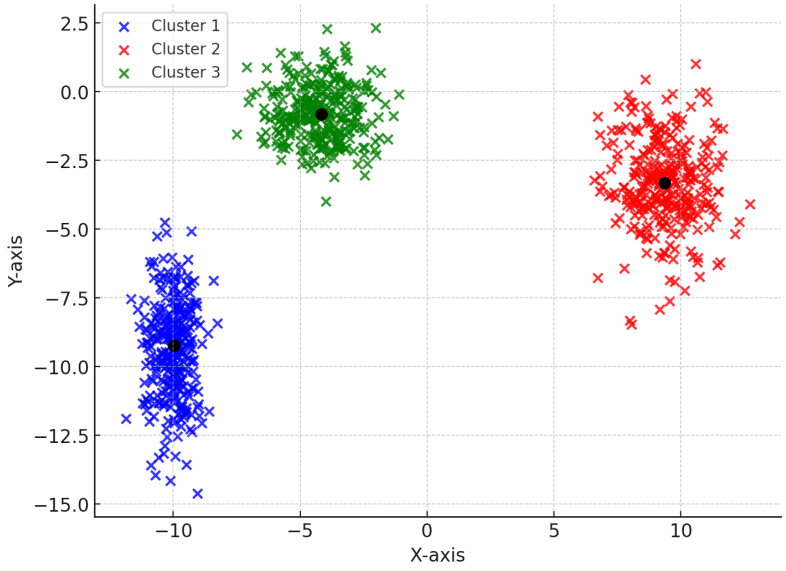

Figure 1: Exemplo de *K-Means*

O *K-Means* é um algoritmo eficiente e rápido, especialmente útil para trabalhar com grandes volumes de dados. Apesar disso, o *K-Means* tem certas limitações. Os resultados dependem fortemente da escolha inicial dos centróides, o que pode levar a soluções inconsistentes. Uma das outras dificuldades, é a necessidade de definir o número de clusters antes de executar o algoritmo, algo que pode ser difícil sem um conhecimento prévio dos dados.

5 - Solução

5.1 Abordagem

Para abordar o tema das *fake reviews*, uma possibilidade é o uso do método *K-Means* que permitirá identificar grupos baseados em características textuais e comportamentais extraídas das avaliações. Esta abordagem é particularmente útil numa fase inicial, uma vez que não precisa de dados rotulados (*labels*), sendo adequada para explorar padrões ocultos nos dados.

Com o progresso do trabalho, e à medida que os resultados forem analisados, será avaliado quais métodos explorar para complementar a nossa abordagem. Esta abordagem permitirá que se tirem conclusões iniciais com base nos agrupamentos gerados e, posteriormente, sejam considerados modelos, mais apropriados, que possam aprofundar a análise. Desta forma, será possível adaptar as técnicas às necessidades específicas do problema e aos desafios que possam surgir.

5.2 Dados Utilizados

Neste estudo, irá ser analisado um dataset compilado pelo *McAuley Lab*, utilizando técnicas de *web scraping* para recolher as avaliações publicamente disponíveis na Amazon. Ele abrange dados desde maio de 1996 até setembro de 2023, incluindo mais de 571,54 milhões de avaliações [7].

Cada observação no dataset representa uma avaliação individual feita por um utilizador para um produto. Estas observações contêm informações textuais, categóricas e numéricas relacionadas com a avaliação e o utilizador. Os dados estão no formato JSON, permitindo fácil manipulação e integração em processos de análise.

5.3 Estrutura do Dataset

No contexto deste estudo, o dataset inclui várias colunas que fornecem informações úteis para a análise das avaliações e dos utilizadores. A tabela 1 descreve as colunas mais relevantes presentes no conjunto de dados.

Table 1: Descrição das colunas do Amazon Reviews Dataset

Coluna	**Descrição**
Rating	Classificação numérica do produto, variando de 1 a 5 estrelas.
Title	Título dado pelo utilizador à avaliação.
Text	Texto detalhado da avaliação.
Images	Links ou dados sobre imagens associadas ao produto e/ou avaliação.
Asin	Identificador único do produto.
user_id	Identificador único do autor da avaliação.
Timestamp	Data e hora do registo da avaliação.
Helpful_Vote	Número de votos que indica que a avaliação foi útil.
Verified_Purchase	Indicador booleano que mostra se a compra foi verificada.

5.4 Categorias de Produtos Selecionadas para Análise

Por razões computacionais, é necessário reduzir o volume dos dados. Logo, optou-se por escolher categorias que mantivessem o equilibro e a qualidade dos dados.

As análises realizadas neste estudo concentrar-se-ão em categorias de produtos cuidadosamente selecionadas, com base na proporção de *reviews* falsas e genuínas, para garantir uma abordagem equilibrada e eficaz na identificação de padrões de comportamento. Essas categorias incluem eletrodomésticos (*appliances*), telemóveis e acessórios (*cell phone and accessories*), roupas, sapatos e joias (*clothing, shoes and jewelry*), produtos de escritório (*office products*), desporto e atividades ao ar livre (*sports and outdoors*) e livros (*books*).

Entre as categorias escolhidas, algumas destacam-se pela elevada proporção de avaliações genuínas. Telemóveis e acessórios, produtos de escritório e livros apresentam as *reviews* mais confiáveis, sendo categorias em que a maioria das avaliações são genuínas. Por exemplo, a categoria de livros tem 81% de avaliações consideradas genuínas, o que faz desta categoria uma excelente referência para identificar padrões típicos de avaliações autênticas [8].

Por outro lado, algumas categorias foram selecionadas devido à maior presença de avaliações falsas, tornando-se úteis para o estudo de padrões fraudulentos. Eletrodomésticos, roupas, sapatos, joias, e desporto e atividades ao ar livre apresentam uma quantidade significativa de *reviews* suspeitas. Estas categorias permitem ao modelo identificar características comuns associadas a *reviews* falsas, como explosões de atividade ou textos genéricos repetitivos [8].

6 - Planeamento

A elaboração do planeamento levou em consideração os prazos estipulados e a complexidade de cada tarefa, permitindo que o trabalho progrida de forma eficiente e organizada.

A planificação é apresentada, visualmente, na figura 2.

- **Definição do Problema e Revisão de Literatura**
  - **Período:** 13/10/2024 a 01/12/2024
  - **Descrição:** Identificação do problema e análise aprofundada da literatura científica para contextualizar o projeto e justificar sua relevância.
- **Pré-Processamento dos Dados**
  - **Período:** 01/11/2024 a 31/12/2024
  - **Descrição:** Coleta e organização dos dados, seguida de limpeza inicial para corrigir inconsistências e garantir a qualidade do conjunto de dados.
- **Análise Exploratória de Dados**
  - **Período:** 01/01/2025 a 31/01/2025
  - **Descrição:** Exploração inicial dos dados para identificar padrões, tendências e possíveis desafios, orientando o desenvolvimento da solução.
- **Desenvolvimento da Solução**
  - **Período:** 01/02/2025 a 15/04/2025
  - **Descrição:** Implementação e otimização de algoritmos ou modelos, incluindo testes e validações preliminares.
- **Análise e Discussão de Resultados**
  - **Período:** 15/04/2025 a 27/06/2025
  - **Descrição:** Avaliação detalhada dos resultados obtidos, comparação com os objetivos iniciais e discussão das implicações.

A figura 2 apresenta a divisão do planeamento através de um gráfico de Gantt.

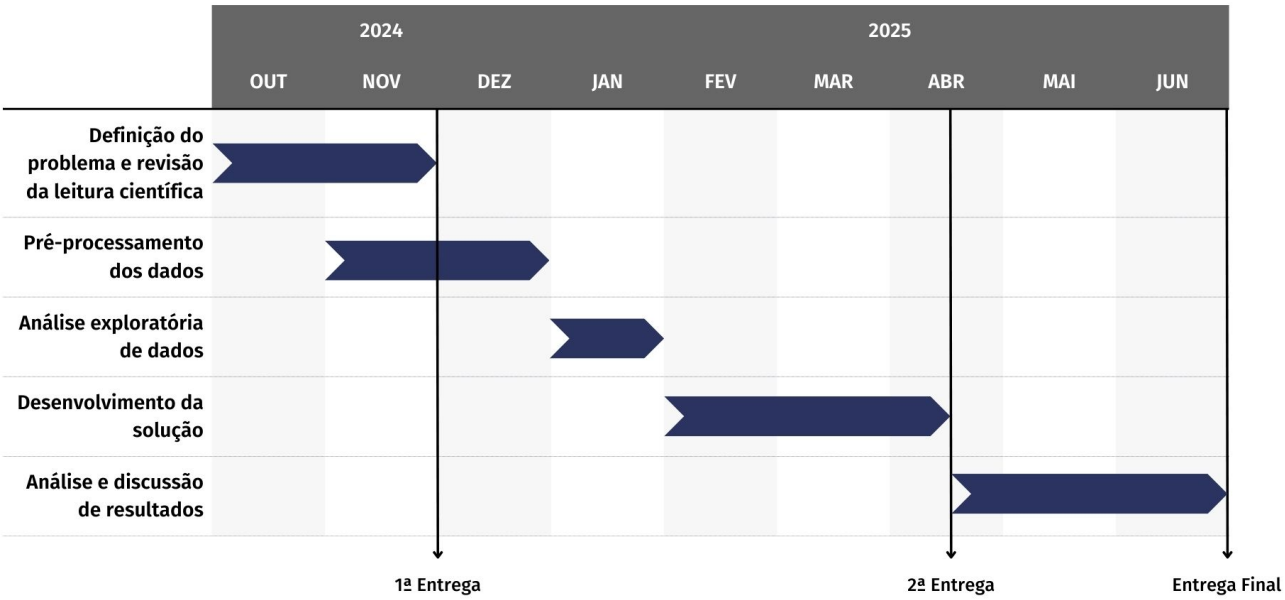

Figure 2: Gráfico *GANTT* do planeamento

Bibliografia

[1] João P. Matos-Carvalho. *The Lusófona LaTeX Template User's Manual*. Lusófona University. 2024. URL: https://github.com/jpmcarvalho/UL-Thesis.

[2] Ramadhani Ally Duma et al. "Fake review detection techniques, issues, and future research directions: a literature review". In: *Knowledge and Information Systems* (2024), pp. 1–42.

[3] Rami Mohawesh et al. "Fake reviews detection: A survey". In: *Ieee Access* 9 (2021), pp. 65771–65802.

[4] A. Mukherjee, B. Liu, and N. Glance. "Spotting Fake Reviewer Groups in Consumer Reviews". In: *Proceedings of the 21st International Conference on World Wide Web (WWW)*. 2012, pp. 191–200.

[5] Bhuvaneshwari P., Rao A. N., and Robinson Y. H. "Spam Review Detection Using Self-Attention Based CNN and Bi-Directional LSTM". In: *Multimedia Tools and Applications* 80.12 (2021), pp. 18107–18124. DOI: 10.1007/s11042-021-10602-y. URL: https://doi.org/10.1007/s11042-021-10602-y.

[6] J. Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv preprint arXiv:1810.04805* (2018). [Online]. URL: http://arxiv.org/abs/1810.04805.

[7] McAuley Lab. *Amazon Reviews 2023*. https://amazon-reviews-2023.github.io/. 2023.

[8] Mozilla. *Fakespot reveals the product categories with the most and least reliable product reviews for summer and back-to-school shopping*. https://blog.mozilla.org/en/mozilla/news/fakespot-reveals-the-product-categories-with-the-most-and-least-reliable-product-reviews-for-summer-and-back-to-school-shopping/. 2024.