UNIVERSIDADE CATÓLICA PORTUGUESA

Machine Learning e Redes Sociais: Automatização da Análise de Comentários para Estratégias de Marketing

Pedro Cunha

Católica Porto Business School
2025

UNIVERSIDADE CATÓLICA PORTUGUESA

Machine Learning e Redes Sociais: Automatização da Análise de Comentários para Estratégias de Marketing

Trabalho Final na modalidade de Dissertação
apresentado à Universidade Católica Portuguesa
para obtenção do grau de mestre em Gestão

por

Pedro Cunha

sob orientação de
Prof. Dr.ª Vera Lúcia Miguéis Oliveira e Silva

Católica Porto Business School, Universidade Católica Portuguesa
Abril 2025

Agradecimentos

Este trabalho não teria sido possível sem o apoio e a contribuição de diversas pessoas, às quais estou profundamente grato.

Em primeiro lugar, gostaria de expressar o meu sincero agradecimento à minha orientadora, Professora Doutora Vera Miguéis, pela excelente orientação que me proporcionou. Obrigado pela disponibilidade, paciência e motivação ao longo de todo este processo.

À empresa que tão bem me acolheu, a Elife Portugal, e destacar a valiosa ajuda dos meus orientadores organizacionais, Margarida Nunes e Rainer. Agradeço também aos meus colegas toda a ajuda proporcionada.

Aos meus pais, que foram o meu alicerce e a minha maior inspiração ao longo destes anos. Obrigado por todo o apoio e confiança inabalável.

À Daniela, pela companhia incondicional em todos os momentos, tanto de conquista como de dúvida. Sem ti, nada disto seria possível.

Aos meus amigos Mota e Barros, pelo apoio constante e pelos momentos de descontração que tanto me ajudaram a enfrentar os desafios desta etapa.

Por fim, a todos os que, de forma direta ou indireta, contribuíram para a realização deste trabalho, o meu mais profundo agradecimento.

Resumo

O crescente volume de dados gerado pelos utilizadores nas redes sociais tem exigido das empresas métodos mais eficientes para a sua análise. Os sistemas tradicionais revelam limitações na interpretação de conteúdo textual, o que justifica a adoção de técnicas de *Machine Learning* (ML) para processar estes dados de forma automatizada. Este estudo testou um sistema de classificação automática de comentários de clientes de uma seguradora nas suas redes sociais, utilizando três algoritmos de ML - *Naïve Bayes* (NB), *K-Nearest Neighbors* (KNN) e *Decision Trees* (DT). Os resultados obtidos indicam que o algoritmo NB alcançou uma taxa de acerto de 65,18% na classificação de sentimentos, enquanto o KNN demonstrou melhor desempenho (85,06% de precisão) na distinção entre conteúdo gerado pela empresa e pelos utilizadores. A categorização de *tags* apresentou maior complexidade, com o NB a atingir 55,28% de taxa de acerto. Os resultados comprovam a viabilidade da aplicação de técnicas de ML na análise automatizada de comentários em redes sociais, oferecendo às empresas uma ferramenta valiosa para monitorizar a perceção dos clientes e orientar estratégias de marketing mais eficazes.

Palavras-chave: Big Data; Comentários; Data Mining; Machine Learning; Previsão; Redes Sociais

Abstract

The increased volume of data generated by users of social networks has required companies to adopt more efficient methods of analysis. Conventional systems have been shown to have limitations in the interpretation of textual content; therefore, there has been an adoption of Machine Learning (ML) techniques to process this data in an automated way. This study evaluated a system that automatically categorises customer comments on an insurance company's social media platforms. The system employed three ML algorithms: Naïve Bayes (NB), K-Nearest Neighbours (KNN) and Decision Trees (DT). The findings demonstrate that the NB algorithm attained 65,18% accuracy in sentiment classification, while KNN exhibited superior performance (85,06% accuracy) in differentiating between content generated by the company and by users. The categorisation of tags proved to be a more complex task, with NB algorithm reaching a hit rate of 55,28%. The results demonstrate the viability of applying ML techniques for the automated analysis of comments on social networks, providing companies with a valuable tool for monitoring customer perceptions and guiding more effective marketing strategies.

Keywords: Big Data; Comments; Data Mining; Machine Learning; Forecasting; Social Networks

Nº palavras: 9032

Índice

Agradecimentos i
Resumo iii
Abstract v
Índice vii
Índice de Figuras ix
Índice de Tabelas xi
Acrónimos xiii

Motivação 1

Capítulo 1: Empresa 3

Capítulo 2: Enquadramento Teórico 4
2.1. Marketing suportado por dados 4
2.2. Redes Sociais e Comportamento do Consumidor 7
2.3. Análise de Sentimentos 10
2.4. Machine Learning 13

Capítulo 3: Metodologia e Dados 17
3.1. Compreensão do Problema 17
3.2. Ferramentas 17
3.2.1. Buzzmonitor 17
3.2.2. RapidMiner 18
3.3. Compreensão dos Dados 18
3.4. Preparação dos Dados 23
3.5. Modelação 24
3.6. Avaliação 25

Capítulo 4: Análise e Discussão dos Resultados 27

Conclusão 40
Declaração de IA generativa e tecnologias assistidas por IA no processo de redação 43
Bibliografia 44

Índice de Figuras

Figura 1: Wordcloud Palavras Total 27
Figura 2: Wordcloud Palavras Positivo 30
Figura 3: Wordcloud Palavras Neutro 30
Figura 4: Wordcloud Palavras Negativo 31
Figura 5: Wordcloud Palavras Earned 33
Figura 6: Wordcloud Palavras Earned 33
Figura 7: Wordcloud Palavras Phishing 36
Figura 8: Wordcloud Palavras Campanhas 36
Figura 9: Wordcloud Palavras Literacia Financeira 37
Figura 10: Wordcloud Palavras Atendimento ao Cliente 37
Figura 11: Wordcloud Palavras Crédito Consolidado 38

Índice de Tabelas

Tabela 1: Origem dos comentários 19
Tabela 2: Sentimento Atual 20
Tabela 3: Origem do Buzz Atual 21
Tabela 4: Tags Atual 22
Tabela 5: Matriz das Tabelas 26
Tabela 6: Matriz da Tabela Tags 26
Tabela 7: Nível de Acerto das Previsões 28
Tabela 8: Sentimento Previsão (NB) 29
Tabela 9: Origem do Buzz Previsão (KNN) 32
Tabela 10: Tags Previsão (NB) 34

Acrónimos

AS – Análise de Sentimento
CE – Customer Experience
CRISP-DM – Cross Industry Process for Data Mining
CRM – Customer Relationship Management
CV – Cross Validation
DT – Decision Tree
IA – Inteligência Artificial
IoT – Internet das Coisas
KNN – K-Nearest Neighbors
ML – Machine Learning
MRS – Marketing de Redes Sociais
NB – Naïve Bayes
PLN – Processamento de Linguagem Natural
RS – Redes Sociais
SAC – Sentiment Analysis Classification
TF-IDF – Term Frequency-Inverse Document Frequency
TM – Text Mining

Motivação

A análise de dados nas Redes Sociais (RS) tem ganho destaque nos últimos anos, com estudos como o de Ghani et al. (2019) que destaca a importância de técnicas avançadas de análise para lidar com o volume massivo de dados gerados por plataformas como Facebook, Twitter e Instagram. A necessidade de técnicas eficientes e abordagens analíticas para lidar com o volume crescente de dados nas RS é cada vez mais urgente (Gama et al., 2014). Estes dados exigem abordagens que vão além dos métodos tradicionais, como sistemas baseados em regras, que são limitados pela sua rigidez e incapacidade de encontrar nuances contextuais.

Na empresa na qual o estudo se baseou, a classificação dos comentários, realizados pelos utilizadores nas RS da empresa, era tradicionalmente realizada através de dois métodos: um sistema de regras e uma classificação manual por parte dos colaboradores. O sistema de regras baseava-se na criação de regras pré-definidas que associam determinadas palavras-chave a classificações específicas. Por exemplo, se um comentário contivesse a palavra "reclamação", era automaticamente classificado como negativo, independentemente do contexto em que a palavra era utilizada. Embora este método permitisse uma certa automatização do processo, apresentava limitações significativas. A principal desvantagem era a sua rigidez, uma vez que como o sistema não considerava o contexto em que as palavras eram utilizadas, muitas vezes classificando os comentários de forma incorreta.

Para superar as limitações, dos métodos tradicionais, e como já referido, é necessário adotar técnicas mais avançadas, como a análise preditiva, que se baseia em métodos estatísticos para descobrir padrões e encontrar relacionamentos complexos nos dados (Sivarajah et al., 2017). Estas abordagens,

como o *Machine Learning* (ML), são essenciais para analisar o contexto e extrair *insights* valiosos de grandes volumes de dados.

O ML, ao contrário do sistema de regras, não se limita a palavras-chave isoladas, mas é capaz de analisar todo o texto, aprendendo a partir de exemplos e identificando padrões complexos nos dados. Este tipo de abordagens permite não só aumentar a precisão das classificações, como também reduzir significativamente o tempo e os recursos necessários para o processo.

Estudos recentes, como o de Mikalef et al. (2018), demonstram que a capacidade de análise de grande volume de dados, incluindo técnicas de ML, é essencial para transformar este fluxo de dados em *insights* valiosos. Estas técnicas permitem às empresas tomar decisões mais informadas e rápidas, melhorando a eficiência operacional e a experiência do cliente, permitindo à empresa obter vantagem competitiva e melhorar o seu desempenho (Mikalef et al., 2018).

A motivação para este estudo vai além da resolução de um problema operacional. Aplicar técnicas de ML na análise de comentários pode trazer benefícios estratégicos para a marca, permitindo uma compreensão mais profunda das necessidades e expectativas dos clientes. Ao automatizar este processo, a empresa poderá reagir de forma mais rápida e eficaz às opiniões dos consumidores, melhorando a sua experiência e fortalecendo a relação com a marca. Este estudo pretende não só contribui para a melhoria dos processos internos da marca, como também abrir portas para a aplicação de soluções semelhantes noutras empresas que enfrentam desafios similares.

Capítulo 1: Empresa

A Elife, fundada em 2004 no Brasil, é uma consultora global especializada em inteligência de mercado e relacionamento multicanal, com presença em vários países, incluindo Portugal, Espanha, México, Colômbia, Chile, Argentina e Peru (Elife, 2025). A empresa dedica-se à análise do comportamento do consumidor através da monitorização das RS e da sua jornada digital, utilizando ferramentas de inovação e tecnologia para aproximar marcas e clientes (Elife, 2025). Com escritórios no Porto desde 2008, a Elife expandiu as suas especializações para áreas como *social media*, *social listening*, *social Customer Relationship Management* (CRM), *customer experience* (CE), *chatbots*, *voicebots*, *social analytics*, *social intelligence*, gestão de crises e *influencer marketing*. Atualmente, a empresa concentra-se em três serviços principais: Inteligência do Consumidor, que analisa a perceção dos consumidores sobre marcas e produtos; *IA – Driven Customer Experience*, que cria e gere *bots* e outros processos de atendimento; e Gestão de CE, que abrange o atendimento multicanal, a cobertura de campanhas promocionais e a gestão de Social CRM.

A Elife destaca-se pela sua paixão em transformar dados em estratégias vencedoras, garantindo que as marcas compreendam e influenciam o processo de decisão dos consumidores de forma eficaz (Elife, 2025).

Capítulo 2: Enquadramento Teórico

O presente capítulo estrutura-se em quatro vertentes principais, articulando conceitos fundamentais para a compreensão da análise de dados na atualidade.

2.1. Marketing suportado por dados

O marketing tem evoluído significativamente ao longo das últimas décadas, refletindo mudanças sociais, económicas e tecnológicas. Inicialmente centrado na promoção de produtos, transformou-se numa prática estratégica orientada para o cliente e, mais recentemente, numa abordagem sustentada pela digitalização e pela utilização intensiva de dados. O Chartered Institute of Marketing (2023) define o marketing como o processo de gestão responsável por identificar, antecipar e satisfazer, de forma lucrativa, os requisitos dos clientes. Esta definição destaca a centralidade dos consumidores e a necessidade de alinhar as estratégias empresariais com os seus desejos, garantindo uma proposta de valor competitiva. Wirtz (2012) complementa esta visão ao descrever o marketing como um plano estratégico competitivo que conjuga atividades funcionais com uma orientação para o cliente, visando a construção de relações comerciais duradouras e lucrativas.

A evolução do marketing reflete a transformação dos paradigmas industriais e tecnológicos. A Revolução Industrial marcou o início deste percurso, ao introduzir máquinas a vapor que aumentaram a eficiência produtiva e reduziram custos (Har et al., 2022). A segunda fase industrial, conhecida como Indústria 2.0, caracterizou-se pelas linhas de produção em massa, permitindo a produção de bens de alta qualidade a preços acessíveis (Har et al., 2022). Com a Indústria 3.0, o avanço da eletrónica, das telecomunicações e da informática fomentou a globalização e o comércio eletrónico, alterando profundamente as dinâmicas de mercado (Har et al., 2022). Por fim, a Indústria 4.0, alavancada pela Internet das

Coisas (IoT) e pela Inteligência Artificial (IA), revolucionou o comportamento dos consumidores e o ambiente empresarial, trazendo consigo novas exigências e oportunidades para o marketing (Har et al., 2022). Neste contexto, surge o conceito de Marketing 4.0, apresentado por Kotler et al., (2016), que reflete a transição de abordagens tradicionais para práticas digitais centradas no consumidor. Este conceito integra quatro pilares fundamentais: identidade da marca, imagem da marca, integridade da marca e interação da marca. Embora os três primeiros já fossem abordados no Marketing 3.0 (Kotler et al., 2019), o Marketing 4.0 destaca a interação da marca como um elemento crucial para criar experiências positivas e estabelecer vínculos emocionais entre os consumidores e as empresas (Morrison & Crane, 2007).

Com o advento da Internet, o marketing enfrentou uma disrupção sem precedentes. As previsões iniciais sugeriam que a Internet poderia reduzir as opções estratégicas das empresas, concentrando-se apenas em táticas de baixo custo (Porter, 2001). Contudo, a realidade mostrou-se oposta. A digitalização gerou novas possibilidades estratégicas, como o marketing de conteúdo, a otimização para motores de busca, o email marketing, a personalização baseada em dados e a publicidade em RS (Ascend2., 2019; Content Marketing Institute, 2025; Key, 2017; Olson et al., 2021). Estas táticas foram organizadas sob a designação abrangente de marketing digital, permitindo às empresas adaptar-se às exigências de um mercado cada vez mais dinâmico e competitivo.

As ferramentas digitais transformaram o marketing num processo mais eficiente e personalizado, atendendo às preferências individuais dos consumidores (Nadler & McGuigan, 2018). Grandhi et al. (2021) aponta que os consumidores modernos são sofisticados e exigentes, resistindo à padronização e preferindo interações personalizadas. Além disso, o comportamento de compra tornou-se mais volátil, exigindo que as empresas compreendam os padrões de

consumo e adaptem as suas estratégias em tempo real (Nadler & McGuigan, 2018).

Na era digital, os dados tornaram-se o núcleo das práticas de marketing. A digitalização e as tecnologias modernas de gestão da informação proporcionam um fluxo constante de informações sobre o comportamento dos consumidores, permitindo às empresas desenvolver estratégias baseadas em evidências (Shah & Murthi, 2021). Esta abordagem, conhecida como *data-driven marketing*, explora os dados recolhidos para identificar padrões de consumo, prever comportamentos futuros e criar campanhas mais eficazes (Rosário & Dias, 2023). Os dados oferecem uma riqueza de informações sobre preferências, hábitos de compra e comportamentos dos consumidores, possibilitando a personalização das ofertas e aumentando a relevância das interações (Rosário & Dias, 2023). Li et al., (2022) destaca que, num ambiente competitivo e incerto, é fundamental que as empresas forneçam produtos e serviços relevantes através de canais adequados e no momento certo. A recolha e análise de dados desempenham um papel crítico neste processo, permitindo uma compreensão profunda dos públicos-alvo e facilitando decisões estratégicas mais informadas (Jiang et al., 2022).

Apesar dos benefícios proporcionados pelo *data-driven marketing*, a sua implementação não está isenta de desafios. A gestão de grandes volumes de dados não estruturados, como *feedbacks* de clientes e publicações em RS, exige infraestruturas robustas e equipas qualificadas (Elias, 2024; Johnson et al., 2021; Suoniemi et al., 2018; Tripathi et al., 2021; Wu et al., 2016). Para superar estes desafios, a análise de dados desempenha um papel crucial, para fornecer informações detalhadas como dados demográficos, padrões de compra e preferências dos consumidores, permitindo uma segmentação mais eficaz (Dam et al., 2019; Jabbar et al., 2020). Além disso, o processamento de dados em tempo real capacita as empresas para responderem rapidamente às mudanças do

mercado e às alterações no comportamento dos consumidores, eliminando atrasos na tomada de decisão e garantindo-lhes uma vantagem competitiva significativa (Yang & Kang, 2017).

O *data-driven marketing* revela-se uma estratégia útil ao recorrer à recolha e análise de dados para otimizar campanhas, tornando-as mais eficazes e personalizadas. Ao compreender melhor os comportamentos e preferências dos consumidores, as empresas conseguem tomar decisões mais informadas e fortalecer a ligação com os clientes através de mensagens mais relevantes (Rosário & Dias, 2023). No entanto, à medida que as RS se tornam uma plataforma central para interação, a digitalização tem demonstrado que estas ferramentas não apenas conectam as marcas aos seus públicos, mas também geram experiências que podem influenciar a satisfação do consumidor, seja a curto ou a longo prazo (Zarantonello & Schmitt, 2010). Assim, ao adotar mecanismos digitais, como as RS, as marcas conseguem obter conhecimentos valiosos que podem ser utilizados para desenvolver estratégias baseadas nas interações com os consumidores, promovendo experiências positivas (DeLorme et al., 2009).

2.2. Redes Sociais e Comportamento do Consumidor

As RS podem ser definidas como ferramentas online que permitem a partilha de conteúdos, opiniões, perspetivas e meios de comunicação (Nair, 2011). Estas têm-se tornado cada vez mais populares devido ao seu acesso conveniente e às opções personalizáveis que oferecem aos utilizadores, criando um ambiente favorável para a interação e participação ativa. Estas plataformas permitem que as marcas e clientes comuniquem diretamente, sem quaisquer restrições de tempo, lugar ou meio, transformando a antiga comunicação unidirecional numa comunicação bidirecional e interativa (Barhemmati & Ahmad, 2015; Kim & Ko, 2012). O crescente número de utilizadores e a sua ampla utilização tornam estas

ferramentas indispensáveis para as empresas, como uma oportunidade para expor bens ou serviços a milhares de pessoas que possam estar envolvidas em comunidades, fóruns, blogs ou RS (Barhemmati & Ahmad, 2015). Em 2023, por exemplo, as estatísticas revelaram que o panorama digital registava cerca de 5,22 mil milhões de utilizadores nas RS, com mais de metade da população global a interagir nestas plataformas, demonstrando o seu alcance exponencial (Statista, 2024).

Com a transição para a era da Web 2.0, a Internet deixou de ser apenas um repositório de informação para se transformar num espaço de interação e colaboração. Esta mudança foi crucial para o surgimento de comunidades virtuais, onde os consumidores passaram a ter um papel mais ativo, partilhando opiniões, experiências e conteúdos de forma dinâmica (Wilson, 2006). O conceito de software social tornou-se uma peça-chave deste novo cenário digital, criando oportunidades para um consumo mais participativo. Neste contexto, o chamado comércio social emergiu como uma forma inovadora de decisão de compra, permitindo que os consumidores se baseiem nas recomendações e experiências de outros utilizadores para fazer escolhas mais informadas (Shin, 2008).

O marketing de redes sociais (MRS) tem vindo a transformar a relação entre empresas e consumidores, tornando a jornada de compra mais interativa e envolvente (Kaplan & Haenlein, 2010; Vithayathil et al., 2020). Através destas plataformas, as marcas conseguem comunicar com milhares de pessoas em simultâneo, ampliando o alcance das suas campanhas e reduzindo custos associados à publicidade tradicional (Ajina, 2019; Ajzen & Fishbein, 1980; Mangold & Faulds, 2009). Mais do que simples canais de divulgação, as RS assumem um papel fundamental na forma como os consumidores tomam decisões, permitindo-lhes aceder facilmente a informações sobre produtos e serviços e impulsionando as compras online (Dwivedi et al., 2021; Pop et al., 2020; Shukla & Nigam, 2018; Vithayathil et al., 2020). Atualmente, muitos

consumidores recorrem a estas plataformas para pesquisar opiniões antes de efetuarem uma compra, tornando-as essenciais na partilha de experiências e recomendações sobre marcas, produtos e serviços (Culnan et al., 2010; Mangold & Faulds, 2009; Mikalef et al., 2013). Pequenos detalhes, como o número de "gostos" num produto no Facebook, podem influenciar significativamente a perceção do consumidor e aumentar a probabilidade de compra ou recomendação (Bhattacharyya & Bose, 2020; Lee et al., 2015). Além disso, fatores como avaliações de outros clientes, recomendações online e a influência de criadores de conteúdo têm demonstrado um impacto significativo na decisão de compra, reforçando o poder das RS na construção da confiança e fidelização do consumidor (Ryu & Park, 2020; Weismueller et al., 2020; Xu et al., 2021).

A eficácia do MRS está também associada a aspetos como a qualidade da informação sobre produtos ou serviços, as experiências emocionais, a confiança na marca, a comunidade e a consciência da marca (Ajina, 2019; Ebrahimi et al., 2021; Hasan & Sohail, 2021). Estudos apontam que elementos como o design das publicações, a confiança nas comunidades das RS, a estrutura das mensagens, as atitudes dos consumidores, as configurações culturais, a experiência com realidade aumentada e a facilidade de compreensão contribuem para moldar o comportamento de compra dos consumidores (Ch et al., 2021; Chang et al., 2020; Chen et al., 2016; Kang & Park, 2018; Kong et al., 2021; Ng, 2013; Whang et al., 2021). Além disso, a responsabilidade social, empatia, raciocínio moral e experiências de utilidade passada, também são capazes de influenciar positivamente a intenção de compra (Rapert et al., 2021). O estudo Worldwide (2008) revela que 70% dos consumidores visitam RS para obter informações sobre produtos, 49% tomam decisões de compra com base nestas informações e 60% dizem provavelmente partilhar os seus conhecimentos com outros consumidores. Neste contexto, a norma subjetiva, ou seja, a influência social percebida pelos consumidores, desempenha um papel moderador crucial, uma

vez que os utilizadores valorizam as opiniões de especialistas e outros compradores, o que contribui para reduzir o impacto do risco percebido e aumentar a confiança, a utilidade percebida e a intenção de compra (Shin, 2008). Contudo, a eficácia desta estratégia depende de um planeamento cuidadoso, que considere os diferentes segmentos demográficos, culturais e comportamentais dos consumidores (Abou-Elgheit, 2018). Sem estratégias bem delineadas e adaptadas às características do público-alvo, as empresas correm o risco de não explorar todo o potencial das RS como ferramenta de marketing (Ryu & Park, 2020; Hanna et al., 2011).

Dado o impacto significativo das RS no comportamento do consumidor e na forma como as decisões de compra são influenciadas, torna-se essencial compreender as opiniões e emoções expressas nas interações online (Bhattacharya et al., 2022). Com milhares de utilizadores a partilharem diariamente experiências e avaliações sobre produtos e serviços, as empresas enfrentam um desafio crescente: interpretar e dar sentido a este vasto fluxo de informação de forma estratégica (Nair, 2011; Rosário & Dias, 2023). No entanto, o volume de dados existente torna inviável a sua análise manual, exigindo o recurso a métodos avançados que permitam extrair padrões e tendências de forma eficiente (Shah & Murthi, 2021). É neste cenário que a Análise de Sentimentos (AS) se destaca, possibilitando a identificação das perceções e emoções dos consumidores através de tecnologias que interpretam a linguagem natural e atribuem significado às interações digitais.

2.3. Análise de Sentimentos

Todos os dias, milhões de pessoas expressam opiniões sobre produtos, marcas e serviços, gerando um fluxo constante de dados textuais (Bhattacharya et al., 2022). No entanto, este volume gigantesco de informação representa um grande desafio para as empresas, tornando praticamente impossível analisá-lo

manualmente. Para compreender o que os consumidores realmente pensam e sentem, as marcas precisam de recorrer a ferramentas avançadas que lhes permitam extrair *insights* estratégicos de forma eficiente (Selimovic et al., 2024).

Para dar sentido a estas interações, as empresas utilizam diferentes abordagens para categorizar as opiniões dos consumidores (Bhattacharya et al., 2022). Métodos baseados em aprendizagem automática e técnicas que recorrem a léxicos predefinidos ajudam a identificar e classificar os sentimentos expressos em comentários e publicações online (Bhattacharya et al., 2022). Desta forma, é possível distinguir entre perceções positivas, negativas ou neutras, proporcionando às marcas uma visão mais clara sobre a forma como os seus produtos e serviços são recebidos pelo público (Bhattacharya et al., 2022). Esta análise não só ajuda a ajustar estratégias de marketing e comunicação, como também permite às empresas responder de forma mais eficaz às necessidades e expectativas dos consumidores (Bhattacharya et al., 2022).

A AS concentrasse no estudo computacional das opiniões e atitudes dos indivíduos relativamente às entidades mencionadas em textos, sendo estas geralmente relacionadas com pessoas ou eventos. Nos últimos anos, abordagens baseadas em redes neuronais têm sido amplamente adotadas, revelando-se eficazes na resolução de diversas tarefas associadas à AS (Song et al., 2019; Xiao et al., 2020; Xiao et al., 2022). Esta análise, combinada com o Processamento de Linguagem Natural (PLN), permitem automatizar a conversão de declarações dos clientes em informações valiosas, fornecendo às empresas *insights* estratégicos de grande relevância (Chen et al., 2022; Pang & Lee, 2008). O PLN, enquanto campo de investigação resultante da interseção entre a ciência da computação e a linguística, tem como principal objetivo capacitar os computadores para compreenderem e processarem a linguagem humana, tanto escrita como falada (Pons, E., 2016). Esta metodologia permite examinar como os consumidores se sentem em relação a determinados produtos ou marcas,

avaliando as emoções e atitudes expressas na linguagem escrita (Liu, 2015). Além da análise de *feedback* do cliente, a AS tem aplicações inovadoras podendo apoiar a tomada de decisão em contextos empresariais, políticos e financeiros (Cui et al., 2023), como a previsão de tendências de mercado e movimentos económicos. Por exemplo, uma pesquisa baseada em dados do Twitter demonstrou que os *insights* obtidos pela AS poderiam ser utilizados para prever movimentos nos mercados de ações (Wang et al., 2020). Este tipo de análise ajuda as empresas a identificar padrões emergentes, ajustar estratégias de marketing e desenvolver novos produtos ou serviços para atender às expectativas em constante mudança dos consumidores (Sweeney et al., 2020).

Ao avaliar os sentimentos expressos nas avaliações dos clientes, nas publicações em RS e em outras interações online, os profissionais de marketing conseguem alcançar uma compreensão mais detalhada de elementos essenciais para a gestão da experiência do cliente. A satisfação do cliente, por exemplo, pode ser avaliada através da identificação de sentimentos positivos em *feedbacks* e avaliações de compras, sendo esta uma ferramenta eficaz para minimizar reclamações e melhorar a perceção da marca. A partilha de experiências e/ou opiniões entre utilizadores quando são positivas, especialmente em relação ao serviço ao cliente, demonstra ter um impacto significativo no comportamento de compra dos consumidores (Rathore & Ilavarasan, 2020). Por outro lado, a análise de sentimentos negativos oferece uma oportunidade para identificar problemas que afetam a experiência dos clientes, permitindo que os profissionais de marketing atuem de forma proativa para eliminar possíveis ameaças e melhorar a relação com o consumidor. Estudos mostram que o tom negativo nas RS pode ter um impacto direto nas vendas de produtos, sublinhando a importância de abordar tais questões rapidamente (Nguyen et al., 2022). A defesa da marca pode ser medida através do sentimento positivo dos clientes mais leais, sendo que a partilha de experiências destes influencia mais as decisões de compra dos

consumidores do que o de outros utilizadores (Duel.tech, 2020). Além disso, possibilitam a identificação de padrões emergentes e mudanças nas preferências dos clientes ao longo do tempo, permitindo às empresas antecipar tendências e ajustar as suas estratégias de forma competitiva (Alsemaree et al., 2024).

2.4. Machine Learning

O processo de AS é um dos domínios onde o ML tem vindo a demonstrar um impacto significativo. O ML desempenha um papel fundamental no *data-driven* marketing, sendo um ramo da IA focado na utilização de algoritmos e modelos estatísticos para identificar padrões em grandes volumes de dados e melhorar continuamente o desempenho dos sistemas informáticos de forma autónoma (Kadyrov & Ignatov, 2019). A capacidade do ML para processar grandes volumes de dados e identificar padrões complexos tem impulsionado a sua aplicação em diversas áreas, incluindo a previsão de riscos de crédito e a otimização de decisões de investimento (Zhao et al., 2025). No domínio da precificação de ativos, algoritmos avançados têm sido utilizados para melhorar a previsão de retornos de ações, permitindo que as empresas tomem decisões mais informadas e maximizem os seus benefícios económicos (Gu et al., 2020; Leippold et al., 2022).

A crescente necessidade de personalização no marketing e o fortalecimento das relações com os consumidores tornaram o ML uma ferramenta indispensável no contexto atual (Drye, 2015). Através da análise de atributos como dados sociodemográficos, padrões comportamentais, perfis psicológicos, interesses e geolocalização, as plataformas de ML possibilitam a criação de estratégias de comunicação altamente personalizadas, promovendo um maior envolvimento do consumidor (Torrens & Tabakovic, 2022; Sundsøy et al., 2014).

A análise preditiva é uma das aplicações mais relevantes do ML no marketing, permitindo antecipar necessidades e preferências dos consumidores e melhorar a experiência do cliente (Moro et al., 2017). Esta abordagem tem sido amplamente

utilizada, possibilitando a identificação de potenciais clientes com maior probabilidade de conversão com base no histórico de compras e nas interações em RS (Drivas et al., 2021; Rezaeian et al., 2021).

Uma grande parte das informações disponíveis nas RS aparece no formato textual e, portanto, não estruturado. Estes dados textuais, como comentários, avaliações e publicações, contêm informações valiosas, mas a sua natureza não estruturada exige técnicas especializadas para extrair informações relevantes, tornando-se o *Text Mining* (TM) essencial para o seu processamento. O TM, também conhecido como mineração de texto, refere-se ao conjunto de técnicas que operam especificamente em dados textuais, permitindo a descoberta de padrões, tendências e associações relevantes (Rajman & Besançon, 1998). Para isso, o TM pode integrar técnicas de PLN, que ajudam a pré-processar os dados textuais, identificando estruturas implícitas, como a gramatical, e transformando textos brutos em informações estruturadas. Estas técnicas são fundamentais para indexar e organizar grandes volumes de dados textuais, seja manualmente ou automaticamente, criando uma base sólida para a descoberta de conhecimento. Por exemplo, a indexação de palavras-chave e a identificação de associações entre termos permitem que os algoritmos de ML processem e analisem textos de forma eficiente, extraindo *insights* que seriam difíceis de obter com métodos tradicionais (Rajman & Besançon, 1998).

Métricas simples, como a contagem de palavras *N-gram* ou a *Term Frequency-Inverse Document Frequency* (TF-IDF), podem ser utilizadas para quantificar a relevância das palavras dentro de um conjunto de documentos analisados (Qaiser & Ali, 2018). O TF-IDF é um método que avalia a importância de uma palavra num documento, calculando o número de vezes que um termo ocorre num documento em relação ao número total de palavras no documento. Por sua vez, os *N-gram* são sequências de n palavras ou *tokens* consecutivos extraídos de um texto, com o objetivo de encontrar padrões e relações contextuais entre

palavras (Ahuja et al., 2019). Estudos demonstram que a aplicação do método TF-IDF melhora o desempenho da AS entre 3% e 4% quando comparado com a utilização do *N-gram* (Ahuja et al., 2019).

No âmbito do ML, a escolha do algoritmo mais adequado desempenha um papel fundamental no desenvolvimento de previsões e classificações. Existem diversos algoritmos que podem ser aplicados em tarefas de previsão e classificação, cada um com as suas características, vantagens e limitações específicas. Conforme Mikalef et al. (2018), a seleção do algoritmo adequado é crucial para maximizar a eficiência e a precisão das previsões, especialmente em cenários de *big data*, onde o volume de dados e a velocidade são fatores críticos. Alguns algoritmos são mais simples e exigem menos recursos computacionais, sendo ideais para conjuntos de dados menores ou quando a velocidade de processamento é uma prioridade.

Tendo em conta o enquadramento teórico, a transição das abordagens tradicionais para práticas baseadas em dados tem sido impulsionada pelo avanço das tecnologias de IA e pela capacidade de análise de grandes volumes de informação, possibilitando estratégias mais personalizadas e eficazes (Rosário & Dias, 2023). No entanto, a gestão e interpretação destes dados representam desafios significativos, como ocorre com os comentários dos clientes nas RS (Rosário & Dias, 2023). A necessidade de ferramentas automatizadas que permitam processar e categorizar estas informações tem-se tornado cada vez mais evidente, uma vez que a análise manual é inviável face ao volume crescente de dados gerado diariamente (Selimovic et al., 2024). Assim, a aplicação de ML no processamento e categorização automática de comentários surge como uma solução eficaz para melhorar a eficiência do marketing digital e a tomada de decisão estratégica das empresas.

Neste contexto, o principal objetivo deste estudo passa por desenvolver um sistema automatizado, baseado em ML, para a categorização de comentários de

clientes, otimizando o processo de análise e interpretação de *feedbacks*. De forma mais específica, pretende-se identificar padrões relevantes nos comentários dos clientes da marca, agrupar automaticamente as opiniões em categorias distintas e avaliar como a automação deste processo pode contribuir para uma gestão mais eficiente da experiência do cliente. Adicionalmente, o estudo visa demonstrar como a aplicação destas metodologias pode facilitar a extração de *insights* estratégicos, permitindo à empresa tomar decisões mais informadas e melhorar a sua comunicação digital.

Capítulo 3: Metodologia e Dados

Para o desenvolvimento deste estudo, foi adotada a abordagem *Cross Industry Process for Data Mining* (CRISP-DM), composta por seis fases: Compreensão do Problema, Ferramentas, Compreensão dos Dados, Preparação dos Dados, Modelação e Avaliação.

3.1. Compreensão do Problema

Para este estudo foram utilizados os dados de um dos clientes de Elife, sendo este uma seguradora que atua em um mercado altamente competitivo. A análise dos comentários dos clientes nas RS é crucial para melhorar a experiência do cliente e otimizar as estratégias de marketing digital. Compreender os comentários permite à empresa identificar pontos fortes e áreas de melhoria, além de responder de forma proativa às necessidades dos clientes.

3.2. Ferramentas

De forma a obter os dados necessários para esta investigação utilizou-se a aplicação Buzzmonitor e o software de ML RapidMiner.

3.2.1. Buzzmonitor

O Buzzmonitor é uma plataforma de monitorização e análise de RS que permite recolher, organizar e interpretar dados de diversas plataformas digitais. Através desta ferramenta, é possível analisar sentimentos, segmentar interações e identificar padrões de comportamento dos utilizadores, ajudando a compreender a perceção do público sobre marcas, produtos e serviços. O Buzzmonitor permite ainda o acesso a funcionalidades como o *social listening* que possibilita a identificação e análise de menções e comentários sobre uma marca e

a gestão do atendimento ao cliente, permitindo às empresas responder de forma mais eficiente às interações dos consumidores. A plataforma disponibiliza ainda *dashboards* interativos e relatórios personalizados, facilitando a visualização dos dados e a tomada de decisões estratégicas (Buzzmonitor, 2025).

3.2.2. RapidMiner

O RapidMiner Studio é uma plataforma de ciência de dados e IA que permite a construção de modelos preditivos sem necessidade de conhecimentos avançados em programação (RapidMiner, 2025). A sua interface visual facilita o desenvolvimento de processos analíticos, permitindo que os utilizadores criem fluxos de trabalho estruturados para análise preditiva (RapidMiner, 2025b).

3.3. Compreensão dos Dados

Os dados utilizados neste estudo foram recolhidos através da plataforma Buzzmonitor, que monitoriza as interações nas RS da marca. A amostra final consistiu em 6.315 comentários, recolhidos entre 1 de julho e 19 de dezembro de 2024. O presente estudo pretende analisar os comentários dos clientes da marca nas RS, com o intuito de, através de um modelo preditivo, classificar os sentimentos (positivo, negativo ou neutro), os comentários em *tags* específicas e a origem de cada comentário.

Para a realização deste estudo, inicialmente foi solicitado acesso aos dados dos comentários publicados nas RS da seguradora para a qual a Elife presta serviços. Durante o processo de extração de dados, foi assegurada a conformidade com as normas de privacidade e proteção de dados. Posteriormente, foi escolhida a aplicação RapidMiner Studio como a plataforma de análise de dados, devido à sua capacidade de realizar análises preditivas e de processamento de texto, sem a necessidade de programação avançada. Ao longo do trabalho foi substituído o

nome original da seguradora por *seguradora* com o objetivo de manter a confidencialidade da marca.

Os dados foram inicialmente classificados da maneira habitual de tratar os dados na empresa, sendo que a sua classificação foi gerada automaticamente pelo Buzzmonitor, utilizando IA e regras pré-definidas para categorizar os conteúdos. Seguidamente, esta análise automatizada foi revista manualmente de modo a garantir uma maior precisão na categorização dos dados.

Os valores obtidos inicialmente, através da plataforma Buzzmonitor, verificaram que a maioria dos comentários foi proveniente do Facebook (74,63%), seguido pelo Instagram (13,15%), Twitter (4,62%), LinkedIn (6,25%), e outras plataformas como blogs, fóruns e menções, que apresentaram valores com menor relevância. A Tabela 1 apresenta a distribuição detalhada dos comentários por plataforma.

Origem Dos Comentários	Amostra Total	
	N	%
Blogs	17	0,27 %
Facebook	4713	74,63 %
Fóruns	11	0,17 %
Instagram	829	13,15 %
Menções	681	10,78 %
Mensagens	29	0,46 %
Insta Story's	119	1,88 %
Linkedin	395	6,25 %
Notícias	58	0,92 %
Twitter	292	4,62 %
Total	6315	100 %

Tabela 1: Origem dos comentários

A predominância de comentários no Facebook reflete a popularidade desta plataforma como canal de interação entre os clientes e a seguradora. O Instagram também se destacou, embora com uma participação menor, indicando que estas duas plataformas são os principais canais de comunicação da seguradora com o seu público.

A análise inicial dos sentimentos revelou que a maioria dos comentários foi classificada como neutra (45,02%), seguida por comentários positivos (27,59%) e negativos (27,40%). A Tabela 2 apresenta a distribuição dos sentimentos nos comentários analisados.

Sentimento	Atual	
	N	%
Positivo	1742	27,59 %
Neutro	2843	45,02 %
Negativo	1730	27,40 %

Tabela 2: Sentimento Atual

A alta percentagem de comentários neutros pode ser explicada pelo fato de muitos comentários serem informativos ou não expressarem uma opinião clara. Por outro lado, a distribuição equilibrada entre comentários positivos e negativos sugere que a empresa enfrenta tanto elogios quanto críticas nas RS. Esta informação é valiosa para a seguradora, pois permite identificar áreas de melhoria e reforçar aspetos positivos da marca.

A análise da origem do *buzz* revelou que a maioria dos comentários foi classificada como *Owned Media* (70,90%), ou seja, publicações feitas nos canais oficiais da seguradora. Os comentários classificados como *Earned Media* (29,10%) representam publicações realizadas por utilizadores fora dos canais oficiais. A Tabela 3 apresenta a distribuição da Origem do *Buzz*.

Origem do Buzz	Atual	
	N	%
Earned Media	1835	29,10 %
Owned Media	4471	70,90 %

Tabela 3: Origem do Buzz Atual

A predominância de *Owned Media* indica que a seguradora tem um controlo significativo da maioria das interações nas RS. No entanto, a presença de *Earned Media* é relevante, pois reflete a opinião espontânea dos consumidores, que pode ser mais autêntica e influenciar outros potenciais clientes.

Os comentários foram categorizados em *tags* específicas, que representam os principais temas abordados pelos clientes. A Tabela 4 apresenta as *Tags* existentes e a sua distribuição percentual.

Tags	Atual	
	N	%
Literacia Financeira	1743	27,71 %
Seguros e Sinistros	747	11,88 %
Outros	724	11,51 %
Phishing	586	9,32 %
Campanhas	350	5,56 %
Crédito Pessoal	293	4,66 %
Concorrência	252	4,01 %
Patrocínios	229	3,64 %
Eventos E Causas	221	3,51 %
Seguro De Saúde	207	3,29 %
Seguradora Pay	193	3,07 %
Imagem Institucional	149	2,37 %
Responsabilidade Social	106	1,69 %
Engagement	80	1,27 %

Crédito Automóvel	71	1,13 %
Crédito Lar	69	1,10 %
Passatempos	54	0,86 %
Atendimento Ao Cliente	49	0,78 %
Segurança Digital	45	0,72 %
Crédito Consolidado	45	0,72 %
Memes	43	0,68 %
RH	16	0,25 %
Juros	12	0,19 %
Cartão De Crédito	4	0,06 %
Crédito Despesas	1	0,02 %
Crédito Férias	1	0,02 %
Crédito Moto	0	0,00 %
Crédito Formação	0	0,00 %

Tabela 4: Tags Atual

A *tag* Literacia Financeira foi a mais mencionada, representando 27,71% dos comentários, o que reflete o interesse dos clientes em iniciativas da seguradora para promover a educação financeira, como conteúdos educativos, workshops e outras ações relacionadas. Outras *tags* relevantes incluem Seguros e Sinistros (11,88%) e *Phishing* (9,32%), que abrange discussões sobre seguros oferecidos pela seguradora (como seguros de vida ou automóvel) assim como relatos de sinistros e processos de reclamação e comentários onde houve tentativas de *phishing* que envolvem a seguradora, incluindo relatos de clientes. A presença de *tags* como Atendimento ao Cliente (0,78%), que inclui *feedback* sobre a qualidade do atendimento ao cliente, como respostas a dúvidas, resolução de problemas e eficiência no serviço e Segurança Digital (0,72%), que aborda a segurança dos serviços digitais da seguradora, incluindo proteção de dados e prevenção de fraudes, sugere que estes temas, embora menos frequentes, ainda são relevantes para os clientes e merecem atenção por parte da empresa.

3.4. Preparação dos Dados

A fase de preparação dos dados foi crucial para garantir a qualidade e a consistência das análises. O pré-processamento dos dados foi realizado utilizando o RapidMiner Studio, uma plataforma de ciência de dados que permite a construção de modelos preditivos sem a necessidade de programação avançada, para transformar os textos brutos em dados estruturados, garantindo que os comentários estivessem prontos para a aplicação dos algoritmos de ML.

O pré-processamento transforma o texto bruto dos comentários em dados estruturados, facilitando a aplicação dos modelos de ML (Rapidminer, 2025c). Para garantir que os dados estejam limpos e organizados, foram aplicadas várias técnicas de pré-processamento, essenciais para melhorar a qualidade da análise.

Uma das primeiras etapas foi a extração de conteúdo, onde o operador identifica e retira blocos de texto relevantes de documentos, evitando a inclusão de palavras isoladas (Rapidminer, 2024). Em seguida, todos os caracteres foram convertidos para minúsculas, garantindo que palavras como "Crédito" e "crédito" sejam tratadas da mesma forma (Rapidminer, 2024b). Outro passo importante passa pela substituição de *tokens*, que permite modificar palavras ou expressões específicas dentro do texto. Esta substituição foi feita com base num dicionário predefinido ou através de expressões regulares, permitindo ajustar o conteúdo conforme necessário antes da tokenização (Rapidminer, 2025d). Aplicou-se a tokenização para dividir o texto em palavras individuais ou frases curtas, separando os elementos mais importantes da mensagem. Dependendo do contexto, é possível definir diferentes regras para a separação, como considerar apenas palavras individuais ou criar pequenas janelas de termos consecutivos (Rapidminer, 2025e). Além disso, foi desenvolvido um conjunto de n-gramas, ou seja, combinações de duas ou mais palavras seguidas, permitindo identificar padrões frequentes na linguagem utilizada pelos clientes (Rapidminer, 2025f). Por fim, utilizou-se dois filtros essenciais: um para remover palavras irrelevantes,

como artigos e preposições que não contribuem para a análise, e outro para eliminar *tokens* com poucos caracteres, garantindo que apenas palavras e expressões significativas sejam mantidas (Rapidminer, 2025g; Rapidminer, 2025h).

3.5. Modelação

Para a construção do modelo preditivo, foram testados três algoritmos de ML: *Decision Tree*, *Naïve Bayes* e *K-Nearest Neighbors*. Cada um destes algoritmos foi escolhido com base na sua capacidade de realizar classificações eficientes com baixo custo computacional. Para a fase de teste dos algoritmos, foi utilizada a técnica de *Cross Validation* (CV), que consiste em dividir o conjunto de dados em k subconjuntos de tamanho igual. Um destes subconjuntos é utilizado como conjunto de teste, enquanto os k - 1 subconjuntos restantes são usados para treinar o modelo. Este processo é repetido k vezes, garantindo que cada subconjunto seja usado exatamente uma vez como conjunto de teste. No final, os resultados das k iterações são combinados para produzir uma estimativa única do desempenho do modelo. A CV é especialmente útil para avaliar a precisão do modelo em dados não vistos e para identificar possíveis casos de *overfitting*, ou seja, quando o modelo se ajusta excessivamente aos dados de treino, mas não generaliza bem para novos dados. O valor de k pode ser ajustado conforme necessário, e, neste estudo, foi utilizado um valor adequado para garantir uma avaliação robusta dos modelos (Rapidminer, 2025i).

- *Decision Tree* (DT): Este algoritmo funciona como um conjunto de regras organizadas em forma de árvore. Cada decisão é baseada em características específicas dos comentários, levando a diferentes caminhos até chegar a uma conclusão. Basicamente, o modelo aprende a classificar os dados dividindo-os em pequenos grupos, tornando a interpretação dos resultados bastante simples e intuitiva (Rapidminer, 2024c).

- *Naïve Bayes* (NB): Um modelo probabilístico que calcula a probabilidade de um comentário pertencer a uma determinada categoria. Este algoritmo parte do pressuposto de que cada palavra no texto contribui de forma independente para a classificação final, o que simplifica bastante os cálculos e permite que o modelo funcione bem mesmo com poucos dados. Além disso, é rápido e eficiente, sendo amplamente utilizado para AS e categorização de texto (Rapidminer, 2024d).
- *K-Nearest Neighbors* (KNN): Classifica os comentários comparando-os com exemplos já conhecidos. O modelo analisa quais são os comentários mais semelhantes e, com base nisso, decide a categoria mais provável. Quanto mais parecido um novo comentário for com exemplos previamente analisados, maior a probabilidade de receber a mesma classificação. Como o KNN depende da proximidade entre os dados, costuma ser necessário normalizar as informações antes de aplicá-lo para garantir que os resultados sejam mais precisos (Rapidminer, 2025j).

3.6. Avaliação

Depois de identificar os padrões, é essencial analisar e compreender os modelos dentro do contexto específico do problema, garantindo que os padrões desenvolvidos sejam relevantes e contribuam para o propósito do estudo (Chawla, 2010). No caso de modelos preditivos, uma das abordagens possíveis é avaliar o nível de acerto do modelo com base nos dados de teste disponíveis (Chawla, 2010).

Foi utilizada uma matriz que quantifica quantos exemplos do *dataset* foram previstos corretamente pelo modelo em análise (diagonal principal), sendo as restantes previsões classificadas de forma incorreta. As colunas correspondem à situação prevista e as linhas à situação atual (cf. Tabela 5). Dada o elevado número de variáveis presentes na categorização por *Tags*, foi criada uma matriz

simplificada que destacasse o nível de exatidão das previsões para cada categoria, focando nas variáveis mais relevantes (cf. Tabela 6).

	Previsão			
Atual	X		Y	
	N	%	N	%
X	Previsão Correta de X	%	Previsão Incorreta de Y (Prevista como X)	%
Y	Previsão Incorreta de X (Prevista como Y)	%	Previsão Correta de Y	%
Total	Previsão Total de X	100 %	Previsão Total de Y	100 %

Tabela 5: Matriz das Tabelas

Tags (Previsão Total p/ Categoria)	Previsão	
Atual	N	%
X (Previsão Total de X)	Previsões Corretas de X	% (Previsões Corretas de X em relação ao número Total de Previsões de X)
Y (Previsão Total de Y)	Previsões Corretas de Y	% (Previsões Corretas de Y em relação ao número Total de Previsões de Y)
Z (Previsão Total de Z)	Previsões Corretas de Z	% (Previsões Corretas de Z em relação ao número Total de Previsões de Z)

Tabela 6: Matriz da Tabela Tags

Capítulo 4: Análise e Discussão dos Resultados

A análise das palavras mais utilizadas nos comentários dos utilizadores (cf. Figura 1) revelou termos que refletem preocupações, experiências e temas recorrentes entre os clientes da seguradora.

Figura 1: Wordcloud Palavras Total

O termo "não" (2086 ocorrências), seguido de "*seguradora*" (1423 ocorrências), o que reforça a centralidade da marca nas discussões, seja para elogiar, criticar ou simplesmente referir-se aos seus serviços. Outras palavras-chave, como "emprestimo" (1085 ocorrências), "credito" (374 ocorrências), "seguro" (277 ocorrências) e "financeira" (277 ocorrências) destacam o foco dos clientes em produtos financeiros, como empréstimos e créditos, assim como reforçam a natureza dos serviços oferecidos pela seguradora. A utilização frequente de "portugal" (1008 ocorrências) e "seguradora_portugal" (493 ocorrências) sugere que muitos comentários estão relacionados ao contexto nacional. Também termos como "enganado" (576 ocorrências), "credores_falsos"

(446 ocorrências), "problemas" (263 ocorrências), "fui_enganado" (298 ocorrências) e "messenger" (388 ocorrências) destacam questões relacionadas com fraudes ou más experiências. Estas palavras-chave revelam não apenas os temas mais utilizados, mas também fornecem *insights* valiosos sobre as principais preocupações e expectativas dos clientes, permitindo à seguradora direcionar as suas estratégias de comunicação e melhorar os serviços prestados de forma mais eficaz.

Após a fase de testes, os algoritmos com melhor desempenho em cada uma das classificações foram selecionados para cada variável a prever. O NB destacou-se na classificação de Sentimentos e *Tags*, enquanto o KNN foi mais eficaz na classificação da Origem do *Buzz*. A Tabela 7 apresenta os resultados obtidos de cada algoritmo, destacando o nível de acerto das previsões.

Nível de Acerto dos Algoritmos	Naïve Bayes	K-Nearest Neighbor	Decision Trees
Sentimento	**65,18 % (+/- 1,45 %)**	64,74 % (+/- 6,60 %)	58,72 % (+/- 2,30 %)
Tags	**55,28 % (+/- 1,47 %)**	36,90 % (+/- 1,39 %)	30,91 % (+/- 4,54 %)
Earned Media/ Owned Media	80,15 % (+/- 1,79 %)	**85,06 % (+/- 0,49 %)**	79,82 % (+/- 1,77 %)

Tabela 7: Nível de Acerto das Previsões

Tendo em conta os resultados obtidos, foi selecionado o NB na classificação de Sentimentos, que obteve um nível de acerto de 65,18%, assim como para a classificação de *Tags*, que obteve 55,28%, enquanto que na classificação da Origem do *Buzz* foi selecionado o KNN que alcançou 85,06%.

Pode-se observar nos resultados da previsão de sentimento realizada pelo modelo (cf. Tabela 8) que se obteve uma taxa de acerto de 78,52% para os comentários classificados como positivos, ou seja, 943 dos 1201 comentários previstos como positivos foram classificados corretamente. Para os comentários

neutros, a taxa de acerto foi de 73,40%, com 1885 acertos em 2568 previsões. Já nos comentários negativos, a taxa de acerto foi de 50,59%, tendo acertado 1288 em 2546 previsões. A tabela revela que o modelo apresentou um maior acerto na previsão das classificações positiva e neutra em comparação com a classificação negativa.

Sentimento	Previsão					
Atual	Positivo		Neutro		Negativo	
	N	%	N	%	N	%
Positivo	943	78,52 %	299	11,64 %	500	19,64 %
Neutro	200	16,65 %	1885	73,40 %	758	29,77 %
Negativo	58	4,83 %	384	14,95 %	1288	50,59 %
Total	1201	100 %	2568	100 %	2546	100 %

Tabela 8: Sentimento Previsão (NB)

Com base na análise das palavras mais utilizadas em cada uma das classificações de Sentimento (cf. Figura 2; Figura 3; Figura 4), observa-se que algumas palavras apresentam um valor de ocorrências bastante significativo e repetem-se entre diferentes classificações, como é o caso de "nao", "portugal", "emprestimo" e "ja", que aparecem tanto em comentários positivos, como neutros e negativos, embora com frequências distintas. Por outro lado, existem palavras que são particularmente relevantes para cada classificação, mostrando uma forte tendência para estarem associadas a um determinado sentimento. Para a classificação positiva, destacaram-se palavras como "*seguradora*", "*seguradora*_portugal", "a_*seguradora*" e "financeira", que reflete uma associação dos comentários classificados como positivos a elogios e temas relacionados à seguradora e a educação financeira. Na classificação neutra, as palavras mais relevantes incluem "iphone", "pagar", "ate" e "descontos",

estando os comentários neutros frequentemente relacionados com campanhas, promoções e informações factuais. Por fim, na classificação negativa, as palavras mais significativas são “enganado”, “emprestimo_sem”, “credores” e “falsos”, indicando que os comentários negativos estão associados a relatos de más experiências, fraudes e insatisfações com produtos ou serviços.

Figura 2: Wordcloud Palavras Positivo

Figura 3: Wordcloud Palavras Neutro

Figura 4: Wordcloud Palavras Negativo

Em relação á classificação de Sentimento, em muitos modelos *Sentiment Analysis Classification* (SAC), revisões neutras não são consideradas (Koppel & Schler, 2005; Koppel & Schler, 2006). Existem duas razões principais:

1. a maioria dos modelos SAC concentram-se na classificação binária, ou seja, na identificação de opiniões positivas vs. negativas;
2. avaliações neutras carecem de informações devido à sua ambiguidade.

No entanto, considerou-se que a classe neutra é fundamental para melhorar o desempenho da classificação de sentimento (Koppel & Schler, 2006). Como a neutralidade é considerada em um lugar entre a positividade e a negatividade, a ideia é tratá-la como ruído potencial do ponto de vista da classificação de filtragem de ruído (Sáez et al., 2016). Entende-se que detetar e remover ruído pode melhorar o desempenho de um modelo. Neutralidade significa a ausência de sentimento ou nenhum sentimento (Liu, 2015). No entanto, considera-se que as opiniões que expressam sentimentos mistos ou conflitantes também podem ser consideradas neutras. As avaliações neutras mostram um peso ambíguo do sentimento, ou seja, contêm uma carga equitativa de polaridade positiva e negativa.

A partir dos resultados da previsão da Origem do *Buzz* (cf. Tabela 9), foi alcançada uma precisão de 81,60% para os comentários categorizados como *Earned Media*. Isso significa que, dos 1413 comentários previstos como *Earned Media*, 1153 foram classificados corretamente. Já para os comentários categorizados como *Owned Media*, a taxa de acerto foi de 86,06%, com 4211 acertos em 4893 previsões. A tabela indica que o modelo demonstrou maior acerto na previsão das classificações *Owned Media* em comparação com as classificações *Earned Media*.

Origem do Buzz	Previsão			
Atual	Earned Media		Owned Media	
	N	%	N	%
Earned Media	1153	81,60 %	682	13,94 %
Owned Media	260	18,40 %	4211	86,06 %
Total	1413	100 %	4893	100 %

Tabela 9: Origem do Buzz Previsão (KNN)

Através da análise das palavras mais frequentes em cada classificação da Origem do *Buzz* (cf. Figura 5; Figura 6), nota-se que algumas palavras possuem um número de ocorrências bastante expressivo e aparecem em ambas as classificações, como é o caso de "seguradora", "portugal", "nao" e "ja". Contudo, existem palavras que aparentam ser relevantes para cada uma das classificações. Nos *Earned Media*, destacam-se termos como "iphone", "credito", "juros" e "pagar", o que sugere que os comentários dessa categoria estão frequentemente ligados a campanhas, promoções e questões financeiras. Já nos *Owned Media*, as palavras mais relevantes incluem "emprestimo", "enganado", "credores" e "falsos", indicando que os comentários dessa classificação estão mais

relacionados a relatos de experiências negativas, fraudes e insatisfações com produtos ou serviços.

Figura 5: Wordcloud Palavras Earned

Figura 6: Wordcloud Palavras Earned

A análise das *tags* revelou padrões distintos na capacidade preditiva do modelo (cf. Tabela 10), com variações significativas nos níveis de acerto entre

categorias. O modelo demonstrou elevada precisão na classificação da *tag* Phishing, alcançando uma taxa de acerto de 90,09%, assim como Campanhas, que apresentou um nível de acerto considerável de 75,36%. A categoria Literacia Financeira, apesar de ter gerado o maior volume de previsões (2091 previsões), registou um nível de acerto de 65,85%. Em contraste, desempenhos mais limitados foram verificados nas categorias Atendimento ao Cliente (14,81%), Crédito Consolidado (7,14%) e Crédito Férias (0,00%), evidenciando desafios específicos na classificação destas *tags*.

Tags (Previsão Total P/ Categoria)	Previsão	
Atual	N	%
Phishing (585)	527	90,09 %
Campanhas (138)	104	75,36 %
Literacia Financeira (2091)	1377	65,85 %
Atendimento ao Cliente (27)	4	14,81 %
Crédito Consolidado (28)	2	7,14 %
Crédito Férias (788)	0	0,00 %

Tabela 10: Tags Previsão (NB)

Para a *tag* Phishing (cf. Figura 7), contou com a presença de termos como "emprestimo" (925 ocorrências), "enganado" (532 ocorrências) e "credores_falsos" (446 ocorrências), cuja natureza descritiva de situações fraudulentas facilitou a sua classificação. No caso das Campanhas (cf. Figura 8), apesar de apresentar uma elevada frequência de termos como "portugal" (132 ocorrências) e "*seguradora*" (96 ocorrências) estes encontram-se entre as

palavras mais frequentes na generalidade dos comentários, porém existem termos bastante relevantes para a sua categorização como é o caso de "volta" (95 ocorrências), "feminina" (88 ocorrências) e "a_portugal" (74 ocorrências), associadas a campanhas específicas realizadas pela seguradora.

A categoria Literacia Financeira (cf. Figura 9), apresentou um nível de acerto moderado, tendo as previsões incorretas sido maioritariamente confundidas com Seguros e Sinistros e Outros. Esta dificuldade decorreu da partilha de termos entre estas 3 categorias como "pagar", "dinheiro", "pessoas" e "saber". Contudo, palavras como "financeira" (166 ocorrências) "descontos" (156 ocorrências), "reforma" (136 ocorrências) e "fidelidade" (95 ocorrências) podem-se considerar relevantes para esta categoria. Também as *tags* de Atendimento ao Cliente e Crédito Consolidado (cf. Figuras 10 e 11), por representarem categorias minoritárias no *dataset* utilizado para treinar o modelo preditivo (cf. Tabela 4), apresentaram um baixo nível de acerto, sendo frequentemente classificadas erradamente como Seguros e Sinistros ou Outros, que representam respetivamente, a segunda e a terceira categoria mais frequente no *dataset*, sugerindo que o desequilíbrio na distribuição das classes comprometeu o desempenho do modelo em categorias minoritárias.

Quanto ao Crédito Férias, a única ocorrência no *dataset* correspondia a um *emoji* que, após pré-processamento texto, o classificado como um comentário em branco. Esta limitação levou o modelo a associar automaticamente comentários sem conteúdo textual à *tag* Crédito Férias (788 previsões), evidenciando a importância da introdução do contexto nos dados de entrada.

Figura 7: Wordcloud Palavras Phishing

Figura 8: Wordcloud Palavras Campanhas

Figura 9: Wordcloud Palavras Literacia Financeira

Figura 10: Wordcloud Palavras Atendimento ao Cliente

Figura 11: Wordcloud Palavras Crédito Consolidado

Relativamente aos resultados obtidos na classificação polinomial (*Tags*) e binária (Origem do *Buzz*), podem ser explicados por vários fatores relacionados com a natureza do modelo e a complexidade da tarefa de classificação. Foi possível notar que os classificadores obtiveram um desempenho superior em classificação binária em comparação com a que apresenta múltiplas classes. Isto sugere que, quando há maior sobreposição entre várias classes, o modelo polinomial encontra dificuldades em generalizar corretamente. Um ponto crítico deve-se ao facto de o modelo polinomial operar de forma a transformar os dados em um espaço de alta dimensionalidade, o que pode levar a um cenário de *overfitting*, especialmente em tarefas mais complexas onde o modelo se adapta demasiado aos dados de treino, mas falham ao generalizar para novos exemplos (Shanableh & Assaleh, 2010). Além disso, ao adicionar interações não lineares mais profundas, pode contribuir para um aumento do risco de o modelo se ajustar ao ruído em vez de identificar padrões reais, o que pode resultar em oscilações bruscas, pela tentativa de prever variações aleatórias nos dados, comprometendo a estabilidade do modelo. Mesmo com técnicas de regularização, o problema de *overfitting* mantém-se, dificultando a generalização

para novos dados. Esta instabilidade pode explicar o desempenho de modelos polinomiais de alta ordem em contextos onde é essencial uma boa capacidade de previsão além do conjunto de treino. (Ivanov & Ailuro, 2024).

Conclusão

Este estudo demonstrou a importância da análise de dados e da mineração de texto (*Text Mining*) na extração de conhecimentos a partir de dados não estruturados, como os comentários nas RS. Através da aplicação de técnicas de ML e da utilização de ferramentas como o RapidMiner, foi possível automatizar a classificação de comentários, identificando padrões de Sentimentos (positivos, neutros e negativos), a sua Origem e categorizando-os em *Tags* específicas. A análise pretende capacitar a empresa em estudo, facultando *insights* valiosos sobre a perceção dos clientes, com vista a melhorar a sua capacidade de resposta e gestão da experiência do cliente. A automatização deste processo não só permite uma resposta mais rápida e eficaz às necessidades dos clientes, como oferecer uma base sólida para uma tomada de decisões estratégica e uma melhor otimização em campanhas de marketing. Estas contribuições são significativas não apenas para a empresa em questão, mas também para outras organizações que enfrentam desafios semelhantes na análise de grandes volumes de dados nas RS.

A metodologia implementada incluiu o pré-processamento dos dados, a aplicação de algoritmos de ML (como o KNN e NB), e a avaliação dos modelos através da CV. Os resultados obtidos revelaram um excelente desempenho na classificação da Origem do *Buzz* (*Earned Media* vs. *Owned Media*), com o algoritmo a alcançar 85,06% de acerto. No entanto, na classificação de sentimentos, que obteve um nível de acerto de 65,18%, assim como na categorização de *Tags*, com um nível de acerto de 55,28%, embora tenham obtido valores relevantes, ficaram aquém do potencial esperado. Apesar disto, estes resultados evidenciam a capacidade de técnicas de ML na análise de grandes volumes de dados, permitindo uma classificação mais rápida e relativamente precisa dos comentários dos clientes, embora com algumas limitações.

Apesar dos avanços alcançados, o estudo enfrentou alguns constrangimentos que podem impactar a precisão dos resultados. Uma das principais limitações está relacionada com a subjetividade inerente à classificação manual dos comentários, utilizada como base para treinar o modelo de ML. A interpretação de um comentário pode variar de pessoa para pessoa, dependendo do contexto, da experiência prévia e da perceção individual de cada avaliador. Sendo o modelo preditivo deste estudo baseado em classificações que foram realizadas tanto por um sistema de regras predefinidas, como manualmente, este fator pode introduzir inconsistências nos dados de treino, o que pode comprometer a fiabilidade do modelo automatizado. Além disso, a forma como os comentários são apresentados na aplicação da empresa representa um desafio adicional. Em muitos casos, um comentário pode ser uma resposta a outro, sendo que a sua correta classificação depende do contexto da interação original. No entanto, a ferramenta utilizada pela empresa não disponibiliza os comentários anteriores em conjunto com os que estão a ser analisados, dificultando a compreensão completa do seu contexto. Isto pode levar a erros na categorização, uma vez que certas expressões ou opiniões podem mudar de sentido quando analisadas isoladamente. Para mitigar este problema, sugere-se que estudos futuros utilizem o histórico de comentários associados a cada comentário, proporcionando um contexto mais amplo para uma classificação mais precisa. Tendo em consideração o conteúdo abordado neste estudo, seria também interessante explorar a integração de algoritmos mais avançados de ML, como redes neuronais profundas, que poderiam melhorar a precisão da classificação, especialmente em cenários com maior complexidade de dados. Uma possibilidade a considerar seria a análise de dados multimodais, combinando texto com imagens, vídeos e outros tipos de conteúdo gerados nas RS, o que permitiria uma compreensão mais abrangente das interações dos clientes. Outra direção promissora poderia passar pela aplicação desta metodologia noutros

setores, como o turismo, o retalho ou a banca, contribuindo para análise da perceção dos clientes nessas áreas e identificar formas de melhorar a experiência do cliente.

Declaração de IA generativa e tecnologias assistidas por IA no processo de redação

Durante a elaboração da minha dissertação, Machine Learning e Redes Sociais: Automatização da Análise de Comentários para Estratégias de Marketing, foram utilizadas as ferramentas ChatGPT e DeepSeek para as tarefas de auxílio no apoio à leitura e síntese de artigos científicos e sugestões de melhoria ao nível da linguagem e pontuação. Apesar de não dispor do registo exato dos *prompts* utilizados (dado que a maioria das interações ocorreram em fases iniciais do processo), declaro que, após a utilização destas ferramentas, revisei e editei todo o conteúdo conforme necessário, assumindo total responsabilidade pelo trabalho final apresentado.

Declaro ainda conhecer e respeitar o Código de Conduta de Inteligência Artificial da Católica Porto Business School.

Bibliografia

Abou-Elgheit, E. (2018). Understanding Egypt's emerging social shoppers. *Middle East Journal of Management, 5*(3), 207-270. https://doi.org/10.1504/MEJM.2018.093611

Ahuja, R., Chug, A., Kohli, S., Gupta, S., & Ahuja, P. (2019). The impact of features extraction on the sentiment analysis. *Procedia Computer Science, 152*, 341-348. http://dx.doi.org/10.1016/j.procs.2019.05.008

Ajina, A. S. (2019). The perceived value of social media marketing: An empirical study of online word-of-mouth in Saudi Arabian context. *Entrepreneurship and Sustainability Issues, 6*(3), 1512. http://doi.org/10.9770/jesi.2019.6.3(32)

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predictiing social behavior*. Englewood Cliffs. https://www.scienceopen.com/book?vid=c20c4174-d8dc-428d-b352-280b05eacdf7

Alsemaree, O., Alam, A. S., Gill, S. S., & Uhlig, S. (2024). Sentiment analysis of Arabic social media texts: A machine learning approach to deciphering customer perceptions. *Heliyon, 10*(9). https://doi.org/10.1016/j.heliyon.2024.e27863

Ascend2. (2019). *2019 digital marketing strategies: Survey summary report*. https://ascend2.com/wp-content/uploads/2018/10/Ascend2-2019-Digital-Marketing-Strategies-Report-181005.pdf

Barhemmati, N., & Ahmad, A. (2015). Effects of social network marketing (SNM) on consumer purchase behavior through customer engagement. *Journal of Advanced Management Science, 3*(4). http://dx.doi.org/10.12720/joams.3.4.307-311

Bhattacharya, S., Sarkar, D., Kole, D. K., & Jana, P. (2022). Recent trends in recommendation systems and sentiment analysis. *Advanced Data Mining*

Tools and Methods for Social Computing, 163-175. https://doi.org/10.1016/B978-0-32-385708-6.00016-3

Bhattacharyya, S., & Bose, I. (2020). S-commerce: Influence of Facebook likes on purchases and recommendations on a linked e-commerce site. *Decision Support Systems, 138*, 113383. https://doi.org/10.1016/j.dss.2020.113383

Buzzmonitor (2025). Buzzmonitor. https://getbuzzmonitor.com/pt/

Ch, T. R., Awan, T. M., Malik, H. A., & Fatima, T. (2021). Unboxing the green box: An empirical assessment of buying behavior of green products. World Journal of Entrepreneurship, *Management and Sustainable Development, 17*(4), 690-710. http://dx.doi.org/10.1108/WJEMSD-12-2020-0169

Chang, H. H., Lu, Y. Y., & Lin, S. C. (2020). An elaboration likelihood model of consumer respond action to facebook second-hand marketplace: Impulsiveness as a moderator. *Information & Management, 57*(2), 103171. https://doi.org/10.1016/j.im.2019.103171

Chartered Institute of Marketing (2023, Março 14). *What is Marketing?* https://www.cim.co.uk/content-hub/quick-read/what-is-marketing/

Chawla, N. V. (2010). Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook*, 875-886. https://doi.org/10.1007/978-0-387-09823-4_45

Chen, J. V., Su, B. C., & Widjaja, A. E. (2016). Facebook C2C social commerce: A study of online impulse buying. *Decision Support Systems, 83*, 57-69. https://doi.org/10.1016/j.dss.2015.12.008

Chen, T., Samaranayake, P., Cen, X., Qi, M., & Lan, Y. C. (2022). The impact of online reviews on consumers' purchasing decisions: Evidence from an eye-tracking study. *Frontiers in Psychology, 13*, 865702. https://doi.org/10.3389/fpsyg.2022.865702

Content Marketing Institute. (2025). *What is content marketing?*. https://contentmarketinginstitute.com/what-%20is-content-marketing/

Cui, J., Wang, Z., Ho, S. B., & Cambria, E. (2023). Survey on sentiment analysis: evolution of research methods and topics. *Artificial Intelligence Review, 56*(8), 8469-8510. https://doi.org/10.1007/s10462-022-10386-z

Culnan, M. J., McHugh, P. J., & Zubillaga, J. I. (2010). How large US companies can use Twitter and other social media to gain business value. *MIS Quarterly Executive, 9*(4), 243-259 https://www.researchgate.net/publication/279893388_How_Large_US_Companies_Can_Use_Twitter_and_Other_Social_Media_to_Gain_Business_Value

Dam, N. A. K., Le Dinh, T., & Menvielle, W. (2019). A systematic literature review of big data adoption in internationalization. *Journal of Marketing Analytics, 7*, 182-195. https://doi.org/10.1057/s41270-019-00054-7

DeLorme, D. E., Huh, J., & Reid, L. N. (2009). Direct-to-consumer advertising skepticism and the use and perceived usefulness of prescription drug information sources. *Health marketing quarterly, 26*(4), 293-314. http://dx.doi.org/10.1080/07359680903304278

Drivas, I. C., Sakas, D. P., Giannakopoulos, G. A., & Kyriaki-Manessi, D. (2021). Search Engines' Visits and Users' Behavior in Websites: Optimization of Users Engagement with the Content. In *Business Intelligence and Modelling: Unified Approach with Simulation and Strategic Modelling in Entrepreneurship 8th* (pp. 31-45). Springer International Publishing. https://doi.org/10.1007/978-3-030-57065-1_3

Drye, T. (2015). Data and creativity: Blessing or curse?. *Journal of Direct, Data and Digital Marketing Practice, 17*(1), 5-7. https://doi.org/10.1057/dddmp.2015.41

Duel.tech. (2020, Janeiro 29). *5 Excellent Examples of Brand Advocacy Programs in 2022*. https://www.duel.tech/blog/excellent-advocacy-program-examples

Dwivedi, Y. K., Ismagilova, E., Hughes, D. L., Carlson, J., Filieri, R., Jacobson, J., ... & Wang, Y. (2021). Setting the future of digital and social media marketing research: Perspectives and research propositions. *International journal of information management, 59*, 102168. https://doi.org/10.1016/j.ijinfomgt.2020.102168

Ebrahimi, P., Khajeheian, D., & Fekete-Farkas, M. (2021). A SEM-NCA approach towards social networks marketing: Evaluating consumers' sustainable purchase behavior with the moderating role of eco-friendly attitude. *International Journal of Environmental Research and Public Health, 18*(24), 13276. https://doi.org/10.3390/ijerph182413276

Elias, A. A. (2024). The 'dark side'of data–driven marketing: A system's thinking analysis. *Journal of Strategic Marketing, 32*(7), 913-929. https://doi.org/10.1080/0965254X.2022.2105741

Elife (2025). Elife. https://elifeportugal.com/sobre/

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM computing surveys (CSUR), 46*(4), 1-37. - https://doi.org/10.1145/2523813

Ghani, N. A., Hamid, S., Hashem, I. A. T., & Ahmed, E. (2019). Social media big data analytics: A survey. *Computers in Human behavior, 101*, 417-428. – https://doi.org/10.1016/j.chb.2018.08.039

Grandhi, B., Patwa, N., & Saleem, K. (2021). Data-driven marketing for growth and profitability. *EuroMed Journal of Business, 16*(4), 381-398. https://www.emerald.com/insight/content/doi/10.1108/emjb-09-2018-0054/full/html

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies, 33*(5), 2223-2273. https://doi.org/10.1093/rfs/hhaa009

Hanna, R., Rohm, A., & Crittenden, V. L. (2011). We're all connected: The power of the social media ecosystem. *Business horizons, 54*(3), 265-273. https://doi.org/10.1016/j.bushor.2011.01.007

Har, L. L., Rashid, U. K., Te Chuan, L., Sen, S. C., & Xia, L. Y. (2022). Revolution of retail industry: from perspective of retail 1.0 to 4.0. *Procedia Computer Science, 200*, 1615-1625. https://doi.org/10.1016/j.procs.2022.01.362

Hasan, M., & Sohail, M. S. (2021). The influence of social media marketing on consumers' purchase decision: investigating the effects of local and nonlocal brands. *Journal of International Consumer Marketing, 33*(3), 350-367. https://doi.org/10.1080/08961530.2020.1795043

Ivanov, A., & Ailuro, S. (2024, March). TMPNN: high-order polynomial regression based on Taylor map factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 11, pp. 12726-12734). https://doi.org/10.1609/aaai.v38i11.29168

Jabbar, A., Akhtar, P., & Dani, S. (2020). Real-time big data processing for instantaneous marketing decisions: A problematization approach. *Industrial Marketing Management, 90*, 558-569. https://doi.org/10.1016/j.indmarman.2019.09.001

Jiang, Z., Li, L., & Wang, D. (2022). MCGM: A multi-channel CTR model with hierarchical gated mechanism for precision marketing. *World Wide Web, 26*(4), 2115-2141. https://doi.org/10.1007/s11280-022-01125-z

Johnson, D. S., Sihi, D., & Muzellec, L. (2021). Implementing big data analytics in marketing departments: Mixing organic and administered approaches to increase data-driven decision making. *In Informatics, 8*(4), 66. https://doi.org/10.3390/informatics8040066

Kadyrov, T., & Ignatov, D. I. (2019). Attribution of customers' actions based on machine learning approach. *CEUR Workshop Proceedings, 2479*, 77-88. https://ceur-ws.org/Vol-2479/project1.pdf

Kang, M. Y., & Park, B. (2018). Sustainable corporate social media marketing based on message structural features: Firm size plays a significant role as a moderator. *Sustainability, 10*(4), 1167. https://doi.org/10.3390/su10041167

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons, 53*(1), 59-68. https://doi.org/10.1016/j.bushor.2009.09.003

Key, T. M. (2017). Domains of digital marketing channels in the sharing economy. *Journal of Marketing Channels, 24*(1-2), 27-38. https://doi.org/10.1080/1046669X.2017.1346977

Kim, A. J., & Ko, E. (2012). Do social media marketing activities enhance customer equity? An empirical study of luxury fashion brand. *Journal of Business research, 65*(10), 1480-1486. https://doi.org/10.1016/j.jbusres.2011.10.014

Kong, H. M., Witmaier, A., & Ko, E. (2021). Sustainability and social media communication: How consumers respond to marketing efforts of luxury and non-luxury fashion brands. *Journal of Business Research, 131*, 640-651. https://doi.org/10.1016/j.jbusres.2020.08.021

Koppel, M., & Schler, J. (2005, July). Using neutral examples for learning polarity. In *International Joint Conference on Artificial Intelligence* (Vol. 19, p. 1616). LAWRENCE ERLBAUM ASSOCIATES LTD. https://www.researchgate.net/publication/220812217_Using_Neutral_Examples_for_Learning_Polarity

Koppel, M., & Schler, J. (2006). The importance of neutral examples for learning sentiment. *Computational intelligence, 22*(2), 100-109. https://doi.org/10.1111/j.1467-8640.2006.00276.x

Kotler, P., Kartajaya, H. and Setiawan, I. (2016) *Marketing 4.0: Moving from Traditional to Digital*. John Wiley & Sons. https://www.drnishikantjha.com/booksCollection/Marketing%204.0_%20

Moving%20from%20Traditional%20to%20Digital%20(%20PDFDrive%20).pdf

Kotler, P., Kartajaya, H., & Setiawan, I. (2019). Marketing 3.0: From products to customers to the human spirit. *Springer Singapore*, 139-156. https://doi.org/10.1007/978-981-10-7724-1_10

Lee, K., Lee, B., & Oh, W. (2015). Thumbs up, sales up? The contingent effect of Facebook likes on sales performance in social commerce. *Journal of Management Information Systems, 32*(4), 109-143. https://doi.org/10.1080/07421222.2015.1138372

Leippold, M., Wang, Q., & Zhou, W. (2022). Machine learning in the Chinese stock market. *Journal of Financial Economics, 145*(2), 64-82. https://doi.org/10.1016/j.jfineco.2021.08.017

Li, K., Dai, G., Xia, Y., Mu, Z., Zhang, G., & Shi, Y. (2022). Green technology investment with data-driven marketing and government subsidy in a platform supply chain. *Sustainability, 14*(7), 3992. https://doi.org/10.3390/su14073992

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139084789

Mangold, W. G., & Faulds, D. J. (2009). Social media: The new hybrid element of the promotion mix. *Business horizons, 52*(4), 357-365. https://doi.org/10.1016/j.bushor.2009.03.002

Mikalef, P., Pappas, I. O., Krogstie, J., & Giannakos, M. (2018). Big data analytics capabilities: a systematic literature review and research agenda. *Information systems and e-business management, 16*, 547-578. https://doi.org/10.1007/s10257-017-0362-y

Mikalef, P., Pateli, A., & Giannakos, M. (2013, Abril). *Why are users of Social Media inclined to Word-of-Mouth?*. Proceedings of the 12th IFIP Conference on e-

Business, e-Services, e-Society. https://doi.org/10.1007/978-3-642-37437-1_10

Moro, S., Cortez, P., & Rita, P. (2017). A framework for increasing the value of predictive data-driven models by enriching problem domain characterization with novel features. *Neural Computing and Applications, 28*, 1515-1523. https://doi.org/10.1007/s00521-015-2157-8

Morrison, S., & Crane, F. G. (2007). Building the service brand by creating and managing an emotional brand experience. *Journal of brand management, 14*, 410-421. http://dx.doi.org/10.1057/palgrave.bm.2550080

Nadler, A., & McGuigan, L. (2018). An impulse to exploit: the behavioral turn in data-driven marketing. *Critical Studies in Media Communication, 35*(2), 151-165. http://dx.doi.org/10.1080/15295036.2017.1387279

Nair, M. (2011). Understanding and measuring the value of social media. *Journal of Corporate Accounting & Finance, 22*(3), 45-51. https://doi.org/10.1002/jcaf.20674

Ng, C. S. P. (2013). Intention to purchase on social commerce websites across cultures: A cross-regional study. *Information & management, 50*(8), 609-620. https://doi.org/10.1016/j.im.2013.08.002

Nguyen, V. T., Tran, T. H. D., & Ngo, T. X. B. (2022). The Influence of Brand Equity on Customer Purchase Decision: A Case Study of Retailers Distribution. *Journal of Distribution Science, 20*(2), 11-18. https://doi.org/10.15722/jds.20.02.202202.11

Olson, E. M., Olson, K. M., Czaplewski, A. J., & Key, T. M. (2021). Business strategy and the management of digital marketing. *Business horizons, 64*(2), 285-293. https://doi.org/10.1016/j.bushor.2020.12.004

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval, 2*(1–2), 1-135. http://dx.doi.org/10.1561/1500000011

Pons, E., Braun, L. M., Hunink, M. M., & Kors, J. A. (2016). *Natural language processing in radiology: a systematic review. Radiology, 279*(2), 329-343. https://doi.org/10.1148/radiol.16142770

Pop, R. A., Săplăcan, Z., & Alt, M. A. (2020). Social media goes green—The impact of social media on green cosmetics purchase motivation and intention. *Information, 11*(9), 447. https://doi.org/10.3390/info11090447

Porter, M. E. (2001). Strategy and the internet. *Harvard Business Review, 79*(3), 62-78. https://hbr.org/2001/03/strategy-and-the-internet

Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications, 181*(1), 25-29. http://dx.doi.org/10.5120/ijca2018917395

Rajman, M., & Besançon, R. (1998). Text mining: natural language techniques and text mining applications. In *Data Mining and Reverse Engineering: Searching for semantics. IFIP TC2 WG2. 6 IFIP Seventh Conference on Database Semantics (DS-7) 7–10 October 1997, Leysin, Switzerland* (pp. 50-64). https://doi.org/10.1007/978-0-387-35300-5_3

Rapert, M. I., Thyroff, A., & Grace, S. C. (2021). The generous consumer: Interpersonal generosity and pro-social dispositions as antecedents to cause-related purchase intentions. *Journal of Business Research, 132*, 838-847. https://doi.org/10.1016/j.jbusres.2020.10.070

Rapidminer (2024). Extract Content. https://docs.rapidminer.com/2024.1/studio/operators/extensions/Web%20Mining/html_processing/extract_html_text_content.html

Rapidminer (2024b). Transform Cases. https://docs.rapidminer.com/2024.1/studio/operators/extensions/Text%20Processing/transformation/transform_cases.html

Rapidminer (2024c). Decision Tree. https://docs.rapidminer.com/2024.1/studio/operators/modeling/predictive/trees/parallel_decision_tree.html

Rapidminer (2024d). Naive Bayes. https://docs.rapidminer.com/2024.1/studio/operators/modeling/predictive/bayesian/naive_bayes.html

Rapidminer (2025). Scalable, Secure Enterprise Data Analytics and AI Platform with Data Fabric. https://altair.com/altair-rapidminer

Rapidminer (2025b). The Design View. https://docs.rapidminer.com/9.7/studio/getting-started/design-view.html

Rapidminer (2025c). Process Documents from Data https://docs.rapidminer.com/10.2/studio/operators/extensions/Text%20Processing/process_document_from_data.html

Rapidminer (2025d). Replace Tokens. https://docs.rapidminer.com/10.0/studio/operators/extensions/Text%20Processing/transformation/replace_tokens.html

Rapidminer (2025e). Tokenize. https://docs.rapidminer.com/latest/studio/operators/extensions/Text%20Processing/tokenization/tokenize.html

Rapidminer (2025f). Generate n-Grams (Terms). https://docs.rapidminer.com/latest/studio/operators/extensions/Text%20Processing/transformation/generate_n_grams_terms.html

Rapidminer (2025g). Filter Stopwords (Dictionary). https://docs.rapidminer.com/latest/studio/operators/extensions/Text%20Processing/filtering/filter_stopwords_dictionary.html

Rapidminer (2025h). Filter Tokens (by Length). https://docs.rapidminer.com/latest/studio/operators/extensions/Text%20Processing/filtering/filter_by_length.html

Rapidminer (2025i). Cross Validation. https://docs.rapidminer.com/9.8/studio/operators/validation/cross_validation.html

Rapidminer (2025j). k-NN. https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/lazy/k_nn.html

Rathore, A. K., & Ilavarasan, P. V. (2020). Pre-and post-launch emotions in new product development: *Insights* from twitter analytics of three products. *International Journal of Information Management, 50*, 111-127. https://doi.org/10.1016/j.ijinfomgt.2019.05.015

Rezaeian, O., Haghighi, S. S., & Shahrabi, J. (2021, Dezembro). *Customer churn prediction using data mining techniques for an Iranian payment application*. 2021 12th International Conference on Information and Knowledge Technology (IKT). https://doi.org/10.1109/IKT54664.2021.9685502

Rosário, A. T., & Dias, J. C. (2023). How has data-driven marketing evolved: Challenges and opportunities with emerging technologies. *International Journal of Information Management Data Insights, 3*(2), 100203. https://doi.org/10.1016/j.jjimei.2023.100203

Ryu, S., & Park, J. (2020). The effects of benefit-driven commitment on usage of social media for shopping and positive word-of-mouth. *Journal of Retailing and Consumer Services, 55*, 102094. https://doi.org/10.1016/j.jretconser.2020.102094

Sáez, J. A., Galar, M., Luengo, J., & Herrera, F. (2016). INFFC: An iterative class noise filter based on the fusion of classifiers with noise sensitivity control. *Information Fusion, 27*, 19-32. - https://doi.org/10.1016/j.inffus.2015.04.002

Selimovic, M., Abd Almisreb, A., & Amanzholova, S. (2024). The Role of Sentiment Analysis in Brand Management and Marketing: A Comparative

Study. *Procedia Computer Science, 251,* 579-584. https://doi.org/10.1016/j.procs.2024.11.152

Shah, D., & Murthi, B. P. S. (2021). Marketing in a data-driven digital world: Implications for the role and scope of marketing. *Journal of Business Research, 125*, 772-779. https://doi.org/10.1016/j.jbusres.2020.06.062

Shanableh, T., & Assaleh, K. (2010). Feature modeling using polynomial classifiers and stepwise regression. *Neurocomputing, 73*(10-12), 1752-1759. - https://doi.org/10.1016/j.neucom.2009.11.045

Shin, D. H. (2008). Understanding purchasing behaviors in a virtual economy: Consumer behavior involving virtual currency in Web 2.0 communities. *Interacting with computers, 20*(4-5), 433-446. https://doi.org/10.1016/S0953-5438(08)00025-8

Shukla, P. S., & Nigam, P. V. (2018). E-shopping using mobile apps and the emerging consumer in the digital age of retail hyper personalization: An insight. *Pacific Business Review International, 10*(10). https://www.researchgate.net/publication/327829218_E-Shopping_using_Mobile_Applications_and_the_Emerging_Consumer_in_the_Digital_Age_of_Retail_Hyper_Personalization_An_Insight

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of business research, 70*, 263-286. – https://doi.org/10.1016/j.jbusres.2016.08.001

Song, Y., Wang, J., Jiang, T., Liu, Z., & Rao, Y. (2019). Targeted sentiment classification with attentional encoder network. In Tetko, I., Kůrková, V., Karpov, P., Theis, F. (Eds.), *Artificial Neural Networks and Machine Learning–ICANN 2019: Text and Time Series* (pp. 93-103). Springer, Cham. https://doi.org/10.1007/978-3-030-30490-4_9

Statista (2024, Outubro). *Number of internet and social media users worldwide as of October*. https://www.statista.com/statistics/617136/digital-population-worldwide/

Sundsøy, P., Bjelland, J., Iqbal, A. M., Pentland, A. S., & de Montjoye, Y. A. (2014, Abril). *Big data-driven marketing: how machine learning outperforms marketers' gut-feeling*. Social Computing, Behavioral-Cultural Modeling and Prediction: 7th International Conference. http://dx.doi.org/10.1007/978-3-319-05579-4_45

Suoniemi, S., Meyer-Waarden, L., & Munzel, A. (2018). Big data-driven marketing: An abstract. *Back to the Future: Using Marketing Basics to Provide Customer Value: Proceedings of the 2017 Academy of Marketing Science (AMS)* (pp. 141-142). Springer. https://doi.org/10.1007/978-3-319- 66023-3_53

Sweeney, J., Payne, A., Frow, P., & Liu, D. (2020). Customer advocacy: A distinctive form of word of mouth. *Journal of Service Research, 23*(2), 139-155. https://doi.org/10.1177/1094670519900541

Torrens, M., & Tabakovic, A. (2022). A banking platform to leverage data driven marketing with machine learning. *Entropy, 24*(3), 347. https://doi.org/10.3390/e24030347

Tripathi, A., Bagga, T., Sharma, S., & Vishnoi, S. K. (2021, Janeiro). *Big data-driven marketing enabled business performance: A conceptual framework of information, strategy and customer lifetime value*. 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence 2011) https://doi.org/10.1109/Confluence51648.2021.9377156

Vithayathil, J., Dadgar, M., & Osiri, J. K. (2020). Social media use and consumer shopping preferences. *International Journal of Information Management, 54*, 102117. https://doi.org/10.1016/j.ijinfomgt.2020.102117

Wang, A., Zhang, Q., Zhao, S., Lu, X., & Peng, Z. (2020). A review-driven customer preference measurement model for product improvement:

sentiment-based importance–performance analysis. *Information Systems and e-Business Management, 18*(1), 61-88. https://doi.org/10.1007/s10257-020-00463-7

Weismueller, J., Harrigan, P., Wang, S., & Soutar, G. N. (2020). Influencer endorsements: How advertising disclosure and source credibility affect consumer purchase intention on social media. *Australasian marketing journal, 28*(4), 160-170. https://doi.org/10.1016/j.ausmj.2020.03.002

Whang, J. B., Song, J. H., Choi, B., & Lee, J. H. (2021). The effect of Augmented Reality on purchase intention of beauty products: The roles of consumers' control. *Journal of Business Research, 133*, 275-284. https://doi.org/10.1016/j.jbusres.2021.04.057

Wilson, J. (2006). 3G to Web 2.0? Can mobile telephony become an architecture of participation?. *convergence, 12*(2), 229-242. https://doi.org/10.1177/1354856506066122

Wirtz, J., Chew, P. Y., Lovelock, C. (2012) *Essentials of Services Marketing* (2nd ed.). Pearson Education. https://www.researchgate.net/publication/273381415_Essentials_of_Services_Marketing_2nd_edition

Worldwide, D. E. I. (2008, Dezembro 5). *Engaging Consumers Online. The impact of social media on purchasing behavior*. https://issuu.com/deiworldwide/docs/dei_study_-_engaging_consumers_online_-_summary

Wu, C. X., Zhao, D. Z., & Pan, X. Y. (2016). Comparison on dynamic cooperation strategies of a three-echelon supply chain involving big data service provider. *Control and Decision, 31*(7), 1169-1177. http://dx.doi.org/10.13195/j.kzyjc.2015.0700

Xiao, L., Hu, X., Chen, Y., Xue, Y., Chen, B., Gu, D., & Tang, B. (2022). Multi-head self-attention based gated graph convolutional networks for aspect-based

sentiment classification. *Multimedia Tools and Applications*, 1-20. https://doi.org/10.1007/s11042-020-10107-0

Xiao, L., Hu, X., Chen, Y., Xue, Y., Gu, D., Chen, B., & Zhang, T. (2020). Targeted sentiment classification based on attentional encoding and graph convolutional networks. *Applied Sciences, 10*(3), 957. https://doi.org/10.3390/app10030957

Xu, X., Zeng, S., & He, Y. (2021). The impact of information disclosure on consumer purchase behavior on sharing economy platform Airbnb. *International Journal of Production Economics, 231*, 107846. https://doi.org/10.1016/j.ijpe.2020.107846

Yang, K.C.C., Kang, Y. (2017, Janeiro 7). *Big Data, Consumer Analytics, and Real-Time Bidding (RTB) Advertising: Emerging International Policy and Regulatory Issues (An Abstract)*. Creating Marketing Magic and Innovative Future Marketing Trends. Developments in Marketing Science: Proceedings of the Academy of Marketing Science. https://doi.org/10.1007/978-3-319-45596-9_99

Zarantonello, L., & Schmitt, B. H. (2010). Using the brand experience scale to profile consumers and predict consumer behaviour. *Journal of brand management, 17*, 532-540. https://doi.org/10.1057/bm.2010.4

Zhao, Y., Xiaogang, B., & Qing-Ping, M. A. (2025). Predicting mergers & acquisitions: A machine learning-based approach. *International Review of Financial Analysis*, 103933. https://doi.org/10.1016/j.irfa.2025.103933