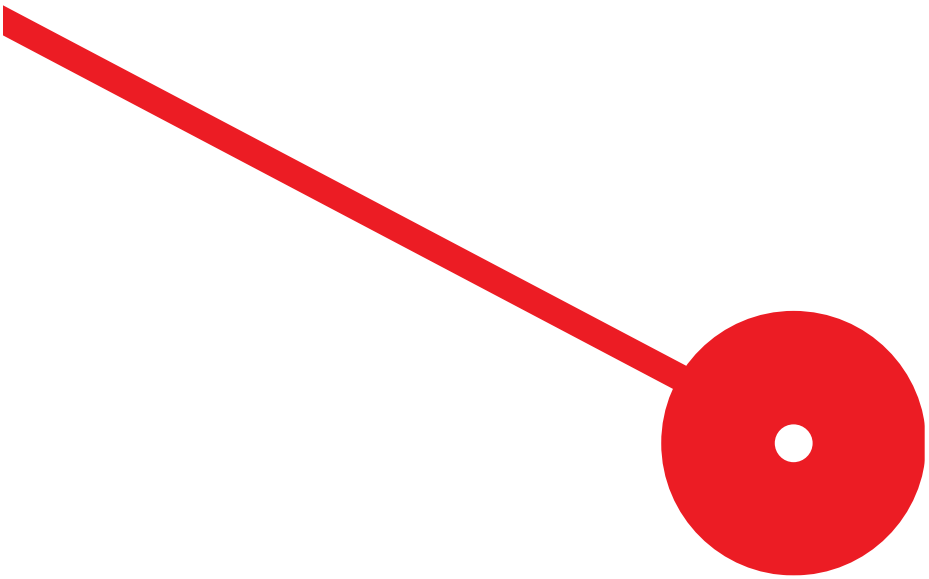
INSTITUTO
SUPERIOR
DE CONTABILIDADE
E ADMINISTRAÇÃO
DO PORTO
POLITÉCNICO
DO PORTO

M MESTRADO
Negócio Eletrónico

Machine Learning como driver do *e-business*

Cláudia Miranda da Costa

10/2022



Cláudia Miranda da Costa . *Machine Learning* como driver do *e-business*
10/2022

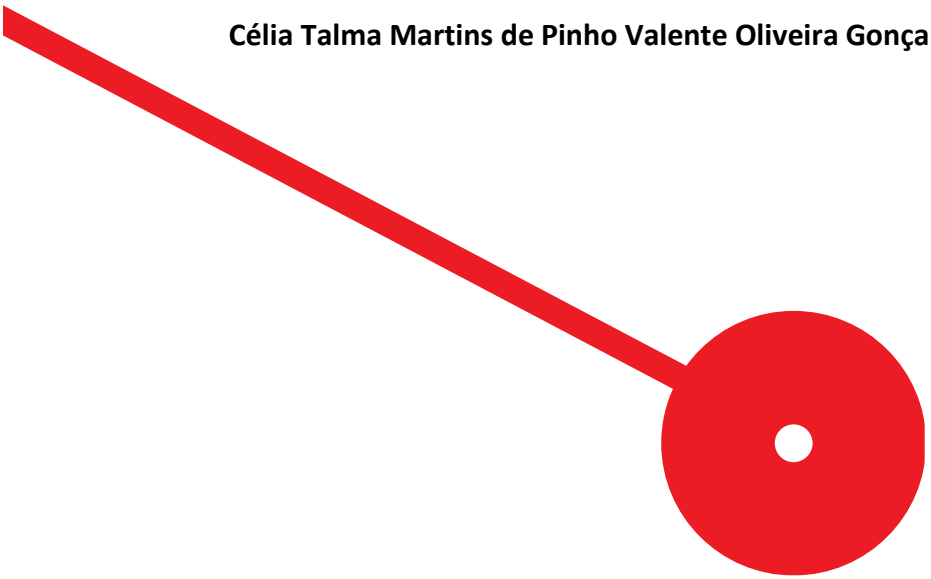
INSTITUTO
SUPERIOR
DE CONTABILIDADE
E ADMINISTRAÇÃO
DO PORTO
POLITÉCNICO
DO PORTO

M MESTRADO
Negócio Eletrónico

Machine Learning como driver do *e-business*

Cláudia Miranda da Costa

Dissertação de Mestrado apresentado ao Instituto Superior de Contabilidade e Administração do Porto para a obtenção do grau de Mestre em Negócio Eletrónico, sob orientação do Professor Doutor Paulo José de Albuquerque Cardoso Trigueiros e Professora Doutora Célia Talma Martins de Pinho Valente Oliveira Gonçalves



Cláudia Miranda da Costa . *Machine Learning* como driver do *e-business*

10/2022

Dedicatória

Dedico esta dissertação a todos os que acreditaram em mim desde o início. Em especial às 4 pessoas mais especiais da minha vida, a minha mãe, o meu pai, o meu irmão e o meu namorado. Estiveram comigo nesta desafiante caminhada. Obrigada!

Agradecimentos

Certamente que as seguintes palavras serão poucas para exprimir o meu agradecimento a quem me acompanhou e contribuiu para a realização desta dissertação.

Deste modo, quero começar por agradecer aos meus orientadores, ao Professor Doutor Paulo Trigueiros e à Professora Doutora Célia Gonçalves, pela enorme paciência, disponibilidade e apoio. Pelos ensinamentos que me transmitiram que, certamente, vou levar para a vida e sem os quais esta dissertação não seria possível. Muito Obrigada!

Aos meus pais, obrigada pela compreensão, pela educação que me proporcionaram ao longo da minha vida.

Ao meu irmão, que nunca me deixou desistir desta etapa, por todos os conselhos e por ser o meu modelo a seguir.

Ao meu namorado, por todo o suporte e ajuda desde o primeiro dia, por acreditar em mim e me dar motivação para enfrentar todos os obstáculos.

Por último, o meu agradecimento a todos os meus colegas de trabalho, pela auxilio e presença, principalmente, durante este último ano, que se mostraram disponíveis para me acompanharem

Muito obrigada! Consegui!

Resumo:

O grande crescimento do comércio eletrónico e o exponencial desenvolvimento tecnológico incentivou à elaboração da presente dissertação. O impacto da inteligência artificial, nomeadamente a aprendizagem máquina que é uma das subáreas da IA, sobre o negócio eletrónico foi a questão inicial desta investigação. A aprendizagem máquina e a forma como é implementada no negócio eletrónico, pretende melhorar o desempenho de experiências e previsões de determinados comportamentos dos consumidores nas compras online, sem estar expressamente programado. No decorrer do estudo foram abordados temas relevantes para a compreensão do impacto da IA no negócio eletrónico tais como: negócio eletrónico e comércio eletrónico, que para uma grande parte da população são conceitos com o mesmo significado, mas, é algo que não se comprova.

O aumento do interesse do público-alvo na partilha de dados (opiniões, crísticas, sentimentos) permitiu-nos recolher um maior volume de informação presente nas plataformas digitais. De modo a evidenciar o impacto da inteligência artificial no negócio eletrónico procedeu-se a uma análise de sentimentos com recurso a algoritmos de aprendizagem máquina. Foram recolhidos *tweets* relacionados com produtos da *Amazon* para posteriormente, utilizarmos técnicas de *text mining* conduzindo à identificação de sentimentos reconhecidos nos *tweets* recolhidos. A informação que conseguimos obter constitui, uma vantagem competitiva para as empresas com negócio eletrónico, saber se o produto foi bem recebido ou não. Desta forma, pretendemos entender qual a posição do atual cliente e de potenciais clientes em relação a determinados produtos ou serviços, com o objetivo de reconhecer o impacto que a inteligência artificial, no negócio eletrónico.

Palavras chave: Negócio eletrónico, Comércio Eletróncio, Inteligência Artificial, *Machine Learning*, *Chatbots*, Sistemas de Recoemndação, *Text Mining*, Análise de Sentimentos.

Abstract:

The procedure of this dissertation was encouraged by the great growth of electronic commerce and the latest exponential technological development. The impact of artificial intelligence – namely machine learning, which is one of the sub-areas of AI – on e-business was the generative fundamental question propelling this investigation. The implementation of machine learning in e-business aims to improve the performance of experiences and predictions of certain consumer behaviors when shopping online – without being expressly programmed. During the course of the study, relevant subjects were addressed to better understand the impact of AI on e-business, such as e-business and e-commerce, that for the majority of people mean the same, which is not correct.

The increasing interest of the generic public in sharing data (e.g. opinions and feelings throughout comments and other interactions) contributed to collecting a greater volume of information present on digital platforms. In order to underline the impact of artificial intelligence on electronic business, sentiment analysis was conducted using machine learning algorithms. Tweets related to Amazon products were collected and processed with text mining techniques leading to the identification of certain feelings. Consequently, we are able to understand the intentions of the current customer, and potential customers, in relation to certain products or services.

Key words: E-business, E-Commerce, Artificial Intelligence, Machine Learning, Chatbots, Recommendation Systems, Text Mining, Sentiment Analysis.

Índice geral

Índice de Figuras

Índice de Tabelas

Índice de Gráficos

Lista de abreviaturas

RCAAP - Repositório Científico de Acesso Aberto de Portugal

IA - Inteligência Artificial

TP - *True Positive*

TN - *True Negative*

FP - *False Positive*

FN - *False Negative*

PLN - Processamento de Linguagem Natural

NER - *Named entity recognition*

SGD - *Stochashe Gradient Descent*

KNN - *K-Nearest neighbors*

DT - *Decision Tree*

Linear SVC - *Linear Support Vector*

Bernoulli NB - *Bernoulli Naive Bayes*

Multinomial NB - *Multinominal Naive Bayes*

RL - Regressão Logistica

TF – *Term Frequency*

IDF - *inverse document frequency*

TF-IDF - *Term frequency–inverse document frequency*

CAPÍTULO - INTRODUÇÃO

A presente dissertação, pretende abordar o impacto da inteligência artificial no negócio eletrónico tendo como base uma análise de sentimentos baseada na extração de informação de redes socias.

O avanço tecnológico continua a criar oportunidades para as empresas. A tecnologia ajuda a melhorar a eficiência, eficácia e qualidade dos serviços fornecidos pelas empresas. A inteligência artificial é uma área em grande desenvolvimento nos nossos dias e que permite criar uma série de desafios e oportunidades para o mercado envolvente. A utilização destas tecnologias leva à criação de sistemas "inteligentes" que podem ajudar a gerir e monitorizar modelos de negócio de forma mais eficaz. (Khrais, 2020)

A inteligência artificial, nem sempre demonstrou a capacidade de satisfazer as exigências do mercado e dos consumidores em diferentes setores. Assim a Inteligência Artificial (IA), aparece ressurgida, recentemente, e está a modificar o panorama económico e a criar mudanças que podem ajudar os consumidores e as organizações a tirar o maior partido devido à utilização destas tecnologias. O negócio eletrónico e nomeadamente o comércio eletrónico é o principal benificiário do aumento da utilização da IA de forma a melhorar os serviços e a experiência das compras online. A IA ajuda a diminuir as complicações que possam surgir de erros humanos, por outro lado pode reduzir as oportunidades de emprego para pessoas que não sejam capazes de utilizar esta tecnologia. De acordo com Sun et al. (Sun et al., 2016)) não podemos deixar de perceber que os benefícios para as organizações são imensos.

Os sistemas dotados de inteligência artificial recolhem e avaliam os dados a um ritmo muito superior quando comparados com os seres humanos. A IA ajuda o comércio eletrónico a captar as tendências de negócio e as necessidades do mercado. A conveniência dos clientes aumenta assim como a satisfação dos mesmos. A utilização desta tecnologia, ajuda por exemplo a melhorar a interação entre as empresas de comércio eletrónico e os seus clientes através da utilização de *chatbots*[1] (Khrais, 2020).

Atualmente, a comunicação entre os consumidores e as organizações acontecem também através da internet, ou seja, redes sociais e sites, daí a importância de entender as opiniões presentes dos consumidores nas redes sociais. É importante acompanhar e perceber o que as pessoas comentam nas redes sociais, para que se possa agir oportunamente.

[1] Programa de computador que simula um ser humano na conversação com as pessoas

a. Motivação

Desde o início, quando me deparei com a possibilidade de vir a defender este tema da dissertação, fiquei logo entusiasmada e muito curiosa. A inteligência artificial é uma área onde eu não tinha nenhum conhecimento prévio, no entanto, sempre me despertou interesse. O comércio eletrónico é em si uma grande tendência no mundo dos negócios, e aliado à Inteligência Artificial pode ajudar a aumentar de forma substancial os resultados das organizações.

Atualmente, informação é sinónimo de poder, significa isto que é cada vez mais importante, para as empresas que detém negócio eletrónico, saber o que é dito ou comentado nas redes socias sobre os bens que são comercializados. Esta informação oferece assim um poder à organização, pois, permite-lhes tomar decisões de forma mais clara e objetiva. É importante conhecer as emoções do consumidor durante o processo de compra dos produtos, se este é bem aceite pelo público-alvo e se vai de encontro às expectativas dos mesmos. Assim, permite mitigar ou até mesmo evitar problemas decorrentes da atividade da empresa com negócio eletrónio

b. Objetivos

O objetivo principal da dissertação é compreender e identificar o impacto da inteligência (IA) artificial no negócio eletrónico, através de uma análise de sentimentos presente em comentários sobre determinado produto. Numa primeira fase, entender conceitos como negócio eletrónico e comércio eletrónico. Investigar a área de conhecimento da inteligência artificial, nomeadamente, a aprendizagem máquina que é uma subárea da inteligência artificial, que está mais diretamente ligada com o negócio eletrónico.

Numa fase posterior, iremos fazer análise de sentimentos, uma tendência recente na análise de textos que tenta identificar as emoções que se encontram por trás dos mesmos.

Desta forma, para dar resposta ao nosso objetivo principal que é o impacto da IA no negócio eletrónico, pretendemos saber o que é dito nas redes socias sobre os produtos da Amazon, conhecer as diferentes emoções que são expressas nos comentários e analisá-las de forma detalhada, através de ferramentas de inteligência artificial

c. Estrutura da dissertação

A estrutura da presente dissertação prende-se numa fase inicial com a revisão da literatura com o objetivo de responder à questão inicial de investigação:

Qual o impacto da inteligência artificial no negócio eletrónico?

Com esta questão pretendemos avaliar de que forma é que a inteligência artificial intervém no negócio eletrónico.

Com a revisão da literatura pretendemos conhecer o estado da arte, ou seja, analisar e sintetizar tudo o que tem sido feito e desenvolvido nesta área de forma a permitir-nos:

- Compreender o negócio eletrónico e comercio eletrónico e as diferenças entre estes dois conceitos;
- Reunir informação sobre a inteligência artificial e a aplicação da mesma no negócio eletrónico;
- Centrar a pesquisa da aprendizagem máquina no *text minig; o* que já foi feito, as conclusões que já foram tiradas, permitindo-nos desta forma fazer a análise de sentimentos.

A nossa investigação tem como ponto de partida a pesquisa em bases de dados, utilizando palavras que foram extraídas a partir da questão inicial de investigação e que posteriormente se foram alargando. As bases de dados utilizadas foram, bibliotecas digitais e sistemas de indexação com artigos científicos tais como, *B-on*, RCAAP, *Google Scholar*, e *Elsevier*. Ao longo da investigação as palavras-chaves e o resumo dos artigos científicos ajudam a identificar os artigos que seriam mais importantes para esta investigação assim como a excluir da investigação.

No segundo capítulo da presente dissertação, com recurso a algoritmos de aprendizagem máquina e ao *text mining*, iremos proceder a uma análise de sentimentos. A análise da presente dissertação foi feita através de *tweets*, neste caso, *tweets* relacionados com o sentimento em relação a determinados produtos da Amazon. Por fim, fizemos a escolha da tecnologia a aplicar, que recaiu sobre a utilização da linguagem *Python*, pelo facto de ser uma linguagem de fácil aprendizagem, e nos permitir criar de forma rápida protótipos para o nosso estudo.

1. Revisão da Literatura

1.1 Negócio Eletrónico e Comércio Eletrónico

O acesso ao comércio eletrónico gerou uma grande mudança social. Diariamente deparamo-nos com a constante e enorme evolução/expansão do comércio eletrónico derivada do desenvolvimento das tecnologias de informação, como por exemplo *smartphones, tablets,* entre outros. (Matana, 2022).

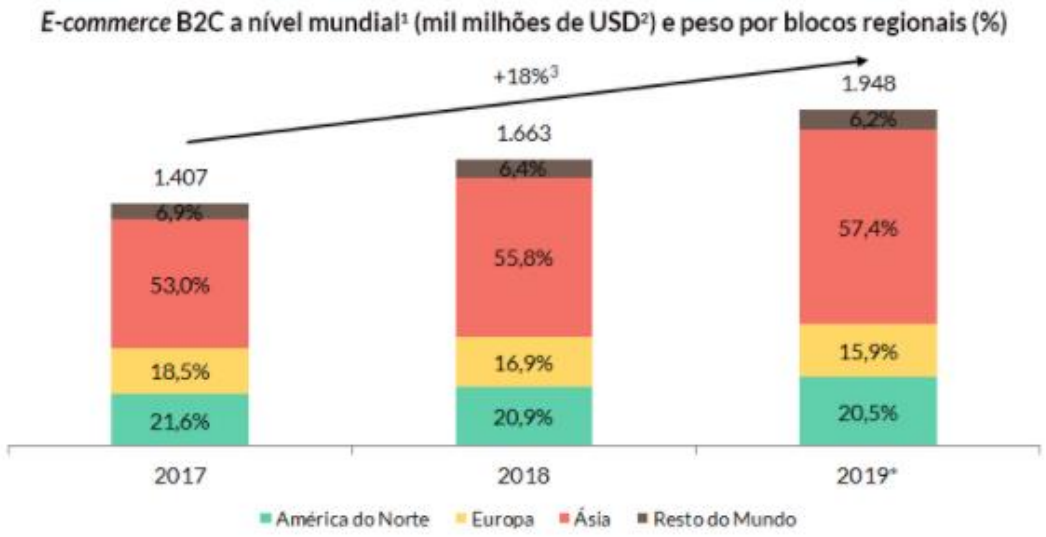


Figura 1- Comércio Eletrónico B2B a nível mundial em *Digital Market Outlook, Statista* janeiro 2020

O negócio eletrónico e comércio eletrónico, são por diversas vezes confundidos. Turban (Turban et al., 2004) define negócio eletrónico tendo em conta um sentido mais vasto do que o comércio eletrónico. O negócio eletrónico caracteriza-se pela transformação das tecnologias da informação e não apenas da transação de bens e serviços como o comércio eletrónico. Tem também como objetivo servir clientes, cooperar com parceiros de negócio e tudo a que diz respeito às transações eletrónicas de uma organização**.**

Como tal, o comércio eletrónico caracteriza-se como a compra e venda de produtos utilizando a internet, bem como pelo uso da tecnologia que consequentemente conduzirá a melhores resultados para as empresas, ou seja, este é definido como o uso da internet para realizar transações de negócio online (DeLone & McLean, 2004).

Para (Turban et al., 2004) o comércio eletrónico *"Refere-se à utilização da Internet e outras redes para comprar, vender, transportar ou comercializar dados, bens ou serviços."*

Existem também outras definições, como a prática de venda de bens e serviços, em que a venda é feita através da internet ou outro sistema, e o pagamento pode ou não ser feito online (Matias, 2016).

O comércio eletrónico é "O uso da Internet, da Web e das aplicações e navegadores que correm em dispositivos móveis para realizar transações de negócio. Mais formalmente, transações comerciais suportadas digitalmente, entre indivíduos e organizações e de organizações entre si" (Laudon & Traver, 2007).

Existem inúmeras vantagens que uma empresa pode retirar da aposta no comércio eletrónico. Atualmente, vivemos num mundo globalizado, e por isso é mais fácil levar o bem ou serviço a qualquer canto do mundo, sendo que ,os custos de pesquisa na internet são bastante reduzidos e por outro lado é mais fácil compará-lo com concorrentes (Laudon & Traver, 2007).

O comércio eletrónico permite que o negócio esteja aberto a qualquer hora, todos os dias da semana, com a desvantagem para algumas indústrias que deixaram de ser intermediárias à medida que os fabricantes construíram relacionamentos diretos com os consumidores (Andonov et al., 2021).

1.2 Inteligência Artificial

A inteligência artificial (IA) é um tema bastante complexo e por isso podemos encontrar diversas definições. A IA, consiste na capacidade das máquinas usarem algoritmos para aprenderem a tomar decisões, a partir de um conjunto de dados e usarem os modelos assim obtidos, tal como um humano o faria (DeLone & McLean, 2004).

As máquinas que funcionam com base em inteligência artificial, não precisam de fazer pausas, não carecem de descanso e ainda tem a capacidade de analisar grandes volumes de dados num certo momento. A proporção de erros é significativamente baixa para as máquinas comparativamente aos humanos.

A IA tem impacto em diversas áreas tais como, a nossa saúde, bem-estar, educação, trabalho e a interação com os outros. Deste modo, a IA oferece-nos sugestões e previsões relativas a questões importantes das nossas vidas, tornando-as assim mais fáceis.

Podemos dizer que a IA tem capacidade de transformar a forma como fazemos negócios, oferecendo uma maior vantagem competitiva para as empresas. Os sistemas que usam inteligência artificial conseguem fazer previsões mais específicas.

Com o uso massivo destas tecnologias, em ambiente empresarial, tem como objetivo promover a eficiência e eficácia das organizações, com a utilização de mecanismos de aperfeiçoamento os processos (Fernandes de Avila et al., 2022).

Deste modo, a IA oferece uma vantagem competitiva, ou seja, com intuito de competir num mercado tão concorrido, como por exemplo auxílio na previsão de comportamento dos consumidores, pois, permite identificar o público-alvo e sugerir ou projetar as mensagens adequadas a cada cliente/consumidor. (DeLone & McLean, 2004)

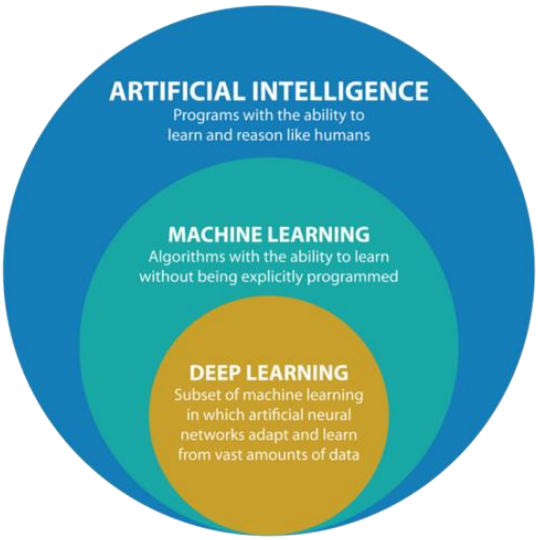


Figura 2- Dimensão da Inteligência Artificial

1.2.1 Aprendizagem Máquina

A aprendizagem máquina (machine learning -ML) progrediu bastante nas últimas décadas. Surgiu como técnica de desenvolvimento de *software* em áreas como a visão por computador, reconhecimento da fala, processamento de língua natural, controle de robôs e outras aplicações.

A aprendizagem máquina é uma subárea da IA e consiste na capacidade de os computadores ou as máquinas aprenderem determinada tarefa sem estarem

explicitamente programadas. Um bom exemplo do que a aprendizagem máquina é capaz de fazer é sugestões ou previsões numa situação particular

"Field of study that gives computers the ability to learn without being explicitly programmed" Arthur Samuel (1959, IBM)

Em 1980, se considerarmos os primeiros computadores pessoais que se tornaram disponíveis para os consumidores, estas máquinas eram, explicitamente, programadas de forma a serem capazes de executar certas tarefas.

Graças à aprendizagem máquina, vários dispositivos que utilizaremos no futuro, terão a capacidade de aprender pela experiência e/ou *feedback* recolhido, de forma a oferecerem experiências personalizadas a cada indivíduo.

A aprendizagem máquina, usa algoritmos específicos que permitem aprender a partir de um conjunto de dados padrão ou através de um conjunto de dados que apresentam determinados padrões. Um exemplo fácil de entender é o uso de filtros do spam no email, que usam ML para detetar e separar o que são efetivamente emails e o que é spam.

Com a aplicação da aprendizagem máquina ao comércio eletrónico obtemos um modelo estatístico, capaz de analisar e classificar milhões de dados sobre o comportamento dos consumidores, e assim conhecer e entender quais os melhores horários da semana para o nosso público-alvo (Dash et al., 2019). Estes sistemas podem aprender e tomar ações/decisões pela análise de um elevado volume de dados. Isto passa pela criação de um modelo , capaz de fazer com que as máquinas consigam prever resultados e comportamentos futuros (Ribeiro & Frazão, 2016).

A aprendizagem máquina divide-se em três áreas: a aprendizagem supervisionada, aprendizagem não supervisionada e a aprendizagem por reforço (Dash et al., 2019).

1.2.1.1 Aprendizagem supervisionada:

Os algoritmos usam informação que já foi filtrada, organizada e pré -classificada. Através deste método, a interação humana é necessária para poder providenciar *feedback.* A aprendizagem supervisionada, que será o foco nesta dissertação, exige um processo prévio de preparação dos dados. Após a preparação e processamento dos dados obtemos um algoritmo, e este está pronto para realizar a tarefa para o qual foi desenvolvido (como por exemplo a análise de sentimentos).

1.2.1.2 Aprendizagem não supervisionada:

Usando algoritmos específicos, são criados modelos a partir de dados não filtrados ou classificados previamente.

Neste tipo de aprendizagem, não existe qualquer tipo de intervenção humana.

1.2.1.3 Aprendizagem por reforço:

Os algoritmos são capazes de aprender através da experiência, na qual os dados gerados são usados para realimentar o sistema e desta forma permitir determinar as próximas ações.

1.2.2 Sistemas de Recomendação

O objetivo de um sistema de recomendação é muito vasto, contundo, para o negócio eletrónico este permite que, a partir da identificação do comportamento do utilizador/consumidor, consiga aumentar o valor do carrinho de compras, gerando assim mais lucros para a organização.

Os métodos mais convencionais passam por sugestões de produtos, tentando assim, persuadir o cliente a comprar outros produtos que estejam relacionados com os últimos artigos visitados. Por exemplo, se o produto em interesse se tratar de um computador, então podemos presumir que o sistema de recomendação irá sugerir algo como um rato.

Com a inteligência artificial, uma empresa consegue recolher e guardar informações e assim avaliar os clientes, garantindo a qualidade dos serviços e produtos. Existe também uma melhoria das interações entre as empresas de comércio eletrónico (*e-commerce)* e os seus clientes, por exemplo através de *chatbots* (Khrais, 2020). Estes podem possuir sistemas alimentados por inteligência artificial ou não, sendo que, se os *chatbots* usarem a aprendizagem máquina podem guardar e lembrar o que o utilizador perguntou em comunicações anteriores e personalizar a conversa atual com base em acontecimentos passados. Quando os *chatbots* são sistemas alimentados por inteligência artificial e aprendizagem máquina proporcionam uma melhor experiência para o utilizador e consequentemente um maior grau de satisfação do cliente para além de uma comunicação mais eficaz.

Os *chatbots* reproduzem o comportamento do ser humano, como por exemplo reconhecer nomes e números em documentos utilizados pelas organizações e simulam o atendimento

feito por uma pessoa. São sistemas de conversação que comunicam com os seres humanos através de língua natural. No quadro seguinte é possível visualizar os dois tipos de *chatbots* (Bertges et al., 2020).

Sistemas baseados em regras pré-definidas	**Sistema alimentos por inteligência artificial e aprendizagem máquina**
Tem regras e respostas pré-definidas na interação.	Retomam conversas de interações, interpretam sentimentos.
Permite escolhas através de botões, são estes que definem as repostas dos utilizadores.	Aprendem enquanto existem, ou seja, aperfeiçoam as respostas com base na experiência e com interações anteriores.
Caso o agente não consiga resolver o problema, este direciona o cliente para uma pessoa humana.	Fazem uso da aprendizagem de máquina, o algoritmo cria modelos, identifica correlações e toma decisões para resolver problemas.
Quando o agente não consegue resolver o problema, deve-se ao fato de não ter entendido o problema ou a pergunta.	Netses sistemas, se o agente não conseguir resolver o problema o cliente é atendido por um humano.

Tabela 1- Tipos de chatbots

Em conjunto com os *chatbots* podemos integrar o processamento de linguagem natural (PLN). Muitos sistemas de aprendizagem máquina permitem a aprendizagem de idiomas e entrada de voz como por exemplo, a Siri ou Alexa. Isto permite que um sistema responda às perguntas dos clientes, resolva problemas e ainda identifique novas oportunidades para vendas.

Estas tecnologias digitais apresentam diversas competências tais como, reconhecimento de imagens e capacidade de tomar decisões (Kaczorowska-Spychalska, 2019) No futuro será cada vez mais recorrente usar mecanismos de processamento de língua natural (M. Mostafa, 2013).

A inteligência humana pode ser por diversas vezes limitada na realização de algumas tarefas no comércio eletrónico, assim, a inteligência artificial vem minimizar a margem de erro e enfrentar desafios crescentes no comércio eletrónico.

Notavelmente, a IA é uma força que está por detrás do sucesso do comércio eletrónico, ajuda a entender as tendências dos negócios e as necessidades do mercado, que se encontra em constante mudança, aumentando assim a conveniência dos clientes.

A inteligência artificial está a modificar o mundo empresarial, que de certa forma tem cada vez mais dificuldade em manter o cliente satisfeito, a utilização da IA é um ponto chave para combater essa mesma dificuldade e aumentar a satisfação dos clientes (Moreira, 2021) .

Os algoritmos de IA realizam tarefas explicitamente definidas com pouca ou até nenhuma intervenção do ser humano, como por exemplo, a transferência de dados de um *e-mail* ou um *call center*[2], para armazenamento numa base de dados, atualizando assim, os dados dos clientes. (Davenport & Ronanki, 2020).

Para Davenport & Ronanki (2020), a IA consegue ganhar *insights*, ou seja, através de grandes volumes de informação sobre o consumidor e as suas transações. Esta informação pode ser de forma numérica, textual, vocal, em forma de imagem ou de expressões faciais.

Na seguinte tabela, mostramos algumas das diferentes aplicações da IA e os respetivos veículos de execução.

Tipos	**Áreas de relacionamento**	**Decisões influencias por IA**
Numérica	Finanças, Contabilidade, vendas	Encomendas, variedade, preço, promoções, investimento
Textual	Satisfação do cliente, análise de avaliação do produto	Modificação do produto, novo produto, melhoria do serviço
Vocal	Apoio ao Cliente	Previsões de compras, atendimento de pedidos

[2] Atendimento telefónico dos clientes

Imagem/vídeo	Análise comportamental do cliente	Conteúdo digital, recomendação de produto

Tabela 2- Tipos de recolhas de insights Fonte: Acedido de "How Artificial Intelligence (AI) is Reshaping Retailing", de V. Shankar, 2018, Journal of

As empresas, que detém negócio eletrónico, podem utilizar e consequentemente beneficiar da IA de diferentes maneiras:

- Perceção e antecipação do comportamento do consumidor.
- Recomendação de produtos, gestão de vendas e relação com o cliente.
- Por outro lado, a nível mais interno da empresa estes podem utilizar a IA para otimizar o inventário, o transporte e as respetivas entregas das encomendas (Shankar, 2018).
- Encaminhar, em tempo real, publicidade adequada a cada cliente de forma a fomentar a vontade de compra do consumidor. (Davenport & Ronanki, 2020)

2. Data Mining

Os dados gerados na internet estão a aumentar rapidamente, seja em grande volume ou em complexidade. Estes mesmo dados podem ser usados para identificar padrões em diversas áreas.

Na imagem seguinte estão presentes as principais características e objetivos do *Data Mining*.

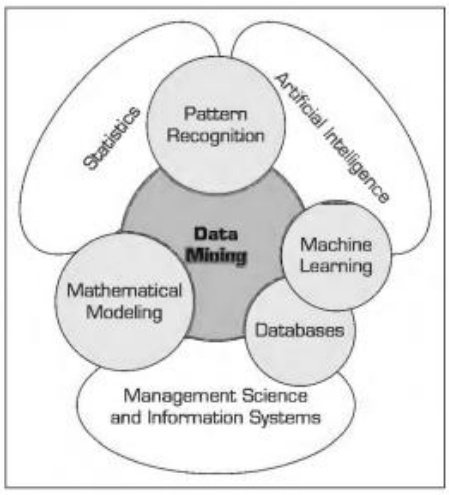


Figura 3- Características Data Mining

Data Mining é amplamente utilizada para melhor direcionar os clientes. Também com o elevado crescimento de dispositivos eletrónicos, o *data mining* torna-se uma ferramenta importante e crucial nos nossos dias.

Assim, *Data Mining* é um termo usado para a descrição e descoberta/exploração de conhecimentos de grandes volumes de dados. O mesmo aparece com outras definições, como por exemplo: extração do conhecimento, análise de padrões, armazenamento de informação e procura de padrões de consumo. O *Data Mining* utiliza diversas disciplinas no seu processo como, estatística, inteligência artificial de forma a extrair e identificar informações uteis e relevantes para organizações. (Turban, 2015)

2.1.1 Aplicações do *Data Mining no negócio eletrónico.*

Gestão do relacionamento com o cliente:

À medida que as empresas constroem relacionamento com os clientes, acumulam enormes quantidades de dados. Quando combinamos estes dados a dados demográficos e atributos socioeconómicos estes podem ser usados para identificar possíveis compradores de novos produtos/serviços, ou seja, permite às organizações entender o perfil do consumidor. O principal objetivo é identificar clientes mais rentáveis permitindo desta forma, maximizar as vendas.

Comércio e logística:

Utilização dos dados para prever volumes de vendas e consequentemente determinar níveis corretos de stock. Identificar possíveis relacionamentos de vendas entre diferentes produtos para melhorar a loja e otimizar as promoções de venda

A indústria de agências de viagem:

Prever vendas de diferentes serviços como por exemplo tipos de quartos em hotéis, tipos de assentos em aviões entre outros de forma a maximizar as receitas e as preferências dos clientes. Identificação de clientes mais rentáveis e fornecimentos de serviços personalizados para manter o seu continuo consumo (Turban, 2015).

2.1.2 Text Mining

O Text Mining pode ser caracterizado como o processo de análise de texto, com o objetivo de extrair informação útil para um determinado fim (Witten, 2004). O *Text minig* apareceu mais tardiamente em comparação com o *Data Minig*.

O *Text Mining* dentro do *Data Mining* é a conversão de texto não estruturado para um estado estruturado, de modo a identificar padrões significativos (Witten, 2004).

O *Text Mining*, tem como principal objetivo, adquirir novos conhecimentos úteis de informação não estruturada (Turban, 2015). Este pode ser aplicado aos mais diversos tipos de documentos com extensões diferentes. Os documentos podem ser HTML, PDF, XML, doc ou até mesmo .txt. Na seguinte figura podemos verificar a estrutura do *text minin*g.

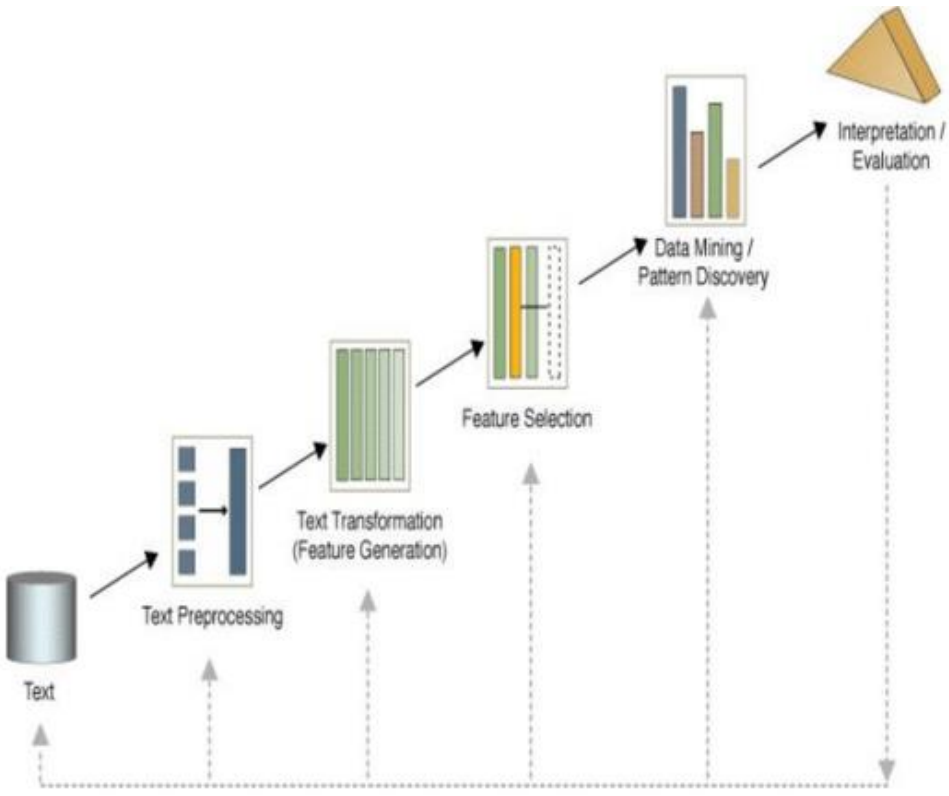


Figura 4 - Text Minig Fonte:(Segall et al., 2022)

Na era da informação em que vivemos, que é caracterizada pelo rápido crescimento na quantidade de dados e informação recolhida, armazenada e disponível no formato eletrónico. A maioria dos dados de negócio armazenada em documentos de texto que não são virtualmente estruturados.

Atualmente, o conhecimento é poder no mundo organizacional e o conhecimento é derivado dos dados e da informação. Com a obtenção de informação relevante ao desenvolvimento da sua atividade, as empresas, terão o conhecimento necessário para tomar melhores decisões, levando assim a uma vantagem competitiva sobre outros organizações.(Turban, 2015)

O *text mining* é definido como um processo semiautomático de extração de padrões de relevância, a partir de grandes quantidades de dados não estruturados (Turban, 2015). O principal objetivo do *Texto mining*, passa pela organização de grandes volumes de dados que se encontram disponíveis, tanto fora como dentro das empresas. Como por exemplo podem ser dados provenientes de *sites*[3] da *web*[4], emprega-os de modo a serem capazes de resolver problemas do mundo real, oferecendo inúmeros proveitos paras as organizações (Godbole et al., 2010).

Para (Turban, 2015) *text mining* consiste no processo de identificar informações válidas, novas, potencialmente úteis, onde os dados são organizados em registos estruturados por variáveis categóricas, ordinais ou contínuas. No processo de *text mining* o conjunto de ficheiros de dados não estruturados (ou semi-estruturados) tais como, documentos de Word, ficheiros PDF, excertos de textos, entre outros. Assim, o *text mining* permite a extração de informações relevantes.

Um exemplo de *text mining* apresentado em (M. Mostafa, 2013) e (Tan et al., 2020) é o desenvolvimento de uma análise de sentimentos para avaliar as opiniões dos consumidores sobre uma câmara digital, através de textos online, extraíram os sentimentos dos consumidores em relação às características interessantes de uma câmara digital como a resolução e a qualidade da fotografia. Deste modo, permite prever futuras vendas de produtos compreendendo os sentimentos dos consumidores.

Os blogs, as avaliações dos produtos em *web sites* é uma *"mina de ouro dos sentimentos dos consumidores"* (Godbole et al., 2010). Este conjunto de informações deve ser corretamente analisado para o aumento da satisfação do cliente. As áreas mais populares da aplicação de *text mining* são:

[3] Páginas web organizadas e localizadas num servidor web
[4] Sistema de informações (infinidade de conteúdos através da internet)

Extração de informação:

Identificação de frases relevantes e relacionamentos dentro de um texto, através da observação de objetos predefinidos e sequências no texto pelo método de combinação de padrões.

Acompanhamento de temas:

Com base no perfil do utilizador e documentos que este visualizou, o *text mining* pode apresentar outras sugestões de documentos análogos que serão de relevância para o utilizador.

Sintetização:

Resumo de texto, para poupar mais tempo ao leitor, ou seja, um único documento, transformando e gerando novo texto.

Categorização:

Identificação dos temas principais de um determinado documento e de seguida colocar os documentos em conjunto predefinido por categorias baseadas nesses temas.

***Clustering*[5]:**

Agrupar documentos semelhantes sem ter um conjunto predefinido de categorias

2.2 Análise de Sentimentos

A Análise de Sentimentos ou mineração de opiniões é um processo automático que deteta o conteúdo emocional ou opinativo que está num determinado texto, sendo que, o texto contém factos ou explicações técnicas sobre um item e também a opinião sobre o mesmo. A partir desta análise conseguimos retirar diversas informações relevantes para a comercialização de um bem ou serviço, tais como, a durabilidade, facilidade de uso de um determinado objeto, perceber se um hotel é recomendado pelos consumidores que já lá estiveram de acordo com as opiniões emitidas. (Cappelli et al., 2016)

Para Liu, 2012 a Análise de Sentimentos, também é chamada de processamento computacional é o campo de estudo que analisa as opiniões dos consumidores,

[5] Ou agrupamento de dados, com base nas semelhanças

sentimentos, avaliações, apreciações, atitudes e as emoções em relação às empresas e aos produtos ou serviços (Liu, 2012).

Nos dias de hoje a análise de sentimentos é relevante, uma vez que diariamente deparamo-nos com grandes volumes de informação, nomeadamente, opiniões, expectativas dos consumidores para com uma marca, A opinião ou o sentimento que retemos sobre algum produto ou serviço é algo importante para praticamente todas as atividades humanas dado que são fundamentais e conseguem influenciar os nossos comportamentos (Liu, 2012). Para as empresas é importante obter as opiniões dos consumidores sobre os produtos e serviços comercializados.

Por outro lado, também é importante para os potenciais clientes conhecerem as opiniões das pessoas que já experimentaram o produto ou serviço, pois estas influenciam na compra. (Liu, 2012). A linguagem, é naturalmente usada para a comunicação entre as pessoas, ou seja, para entendermo-nos uns aos outros.

O enorme crescimento da indústria digital e consequentemente o aumento de plataformas de discussão, comércio eletrónico, *sites* de avaliação de produtos, os média, a rede social entre outras plataformas facilita muito o aumento e o fluxo continuo de pensamentos e opiniões sobre determinado produto ou serviço. Com este aumento de informação, troca de pensamentos e opiniões torna-se um grande desafio para as empresas entender melhor os clientes e potencias clientes, de modo a perceber as suas atitudes e opiniões.(Demircan et al., 2021)

Atualmente, com o crescimento e a importância das redes sociais, as organizações estão cada vez mais a usar este conteúdo para a tomada de decisão e consequentemente para gerar valor para a empresa (Liu, 2012). Se considerarmos um texto subjetivo, em que este representa a avaliação de um utilizador sobre um determinado produto, é de enorme importância a empresa conseguir determinar o que o cliente acha acerca do produto, para deste modo conseguir responder às expectativas do cliente.

Existem duas abordagens fundamentais de análise de sentimentos, os métodos de aprendizagem supervisionada e uma abordagem baseada em léxico. Muitos sistemas de deteção de sentimentos usam métodos baseados em léxico em que constituem uma lista de palavras e a emoção que elas transmitem. Uma desvantagem da utilização de léxicos consiste em as pessoas que transmitem/expressam emoções de diferentes maneiras no

texto. A aprendizagem supervisionada é uma subcategoria da aprendizagem máquina em que os dados são rotulados (Demircan et al., 2021).

Na análise de sentimentos podemos ter essencialmente três categorias: sentimento positivo, sentimento negativo e sentimento neutro. Os sentimentos presentes em avaliação no online afetam diretamente a confiabilidade percebida entre os consumidores. As avalizações negativas têm uma grande influencia na confiança do que as avaliações positivas.

Em diversas plataformas de comércio eletrónico as avaliações são expostas com base em classificação de sentimentos, como por exemplo, a *Amazon* agrupa explicitamente as avaliações dos clientes em categorias favoráveis (sentimento positivo) e críticas (sentimento negativo) (Nakayama & Wan, 2018).

2.2.1 Pré- Processamento

Após a recolha dos dados é necessário proceder ao pré-processamento de texto.

Os dados a utilizar em *texto mining* devem passar por várias transformações, ou seja, pré-processamento que será utilizado como input na construção de um modelo. Esta tarefa de pré-processamento é conduzida pelo processamento de linguagem natural (Turban, 2015). Processamento de Linguagem Natural PLN tem vários níveis de conceitos linguísticos, tais como:

Morfológico, está relacionado com tratamento das palavras;

Léxico, referente ao significado das palavras e do *part-of-speech*[6];

Sintático, que se refere à estrutura da frase e trabalha a gramática;

Fonético, que trata da pronúncia;

Semântico, reflete o significado das palavras, frases e discursos;

Pragmático, neste retrata o conhecimento presente nas pessoas (Feldman, 1999).

Nem todas estas áreas são aplicadas ao *text mining*. Para o pré-processamento textual o primeiro passo para a construção de um documento é selecionar uma base de dados com

[6] Classe gramatical particular da palavra

textos avaliativos, ou seja, texto que possua um sentimento em relação a um produto (Cappelli et al., 2016).

2.2.1.1 Enrichment/Tagging:

Adicionar informação extra às palavras (Gouveia, 2019). Com a aplicação deste processo, dá-se a criação de um tipo de dados denominado de termo. A cada palavra é adicionado um *tag*[7] onde contém informação variada sobre a mesma.

Existem vários *taggers* tudo depende da informação que pretendemos adicionar. Como por exemplo, reconhecimento de uma entidade ou pessoa através *Named entity recognition (NER*), em que o algoritmo identifica se a palavra faz parte de uma entidade ou pessoa.

2.2.1.2 *Stemming*[8]*:*

Consiste no processo de apagar sufixos das palavras para recuperar os radicais. Deste modo, reduz a complexidade sem qualquer perda grave de informação.(Feinerer, 2008).

2.2.1.3 Remoção de *Stopwords*[9]**:**

Ao filtrar os caracteres com a remoção de palavras que na maioria das situações ou contextos são irrelevantes, ou seja, palavras não nos permite fazer a classificação do texto (categorização) como por exemplo (o,a,para,é,onde entre outros) (Liu, 2012).

2.2.1.4 Detecção de Sinóminos

Sinónimos: Em diversos momentos é vantajoso conhecer sinónimos para uma determinada palavra.

2.2.1.5 *Part of Speech Tagging*:

É o processo de marcação das palavras num texto que corresponde a um determinado painel como, substantivos, verbos, adjetivos, advérbios entre outros. Com base na definição da palavra e do contexto em que esta é usada (Turban, 2015, p. 2).

2.2.1.6 Exemplos de aplicação de Análise de Sentimentos nas Redes Sociais

[7] Etiqueta rótulo
[8] Processo de redução de palavras derivadas à sua raiz
[9] Palavras consideradas irrelevantes

Existem diversos estudos que analisam os padrões de conversação nos grupos com a finalidade de uma melhor classificação de sentimento e mais precisas.

Num estudo analisado o foco foi nos comentários do Facebook, reações e partilhas. Deste modo, procederam à extração das publicações, cometários entre outras interações num determinado período. As conclusões retiradas foram que os utilizadores estão mais propensos a clicar em gosto, comentar, partilhar e reagir se concordarem com o que acabaram de ler. Uma publicação receberá mais comentários se os utilizadores concordarem ou discordarem fortemente da referida publicação. Da mesma forma, o utilizador só partilha uma publicação se este sentir que a informação vale a pena ser lida por um público mais amplo.(Kaur et al., 2019).

Em outro estudo, utilizaram o Twitter, nomeadamente tweets, para proceder à análise de sentimentos dos consumidores para com as marcas Nokia, T-Mobile, IBM, KLM, Lufthansa, DHL. A análise de sentimentos neste estudo permitiu perceber que a maioria dos comentários dos tweets foram positivos para a Lufthansa e DHL. Por outro lado, a maioria dos tweets negativos foram para a T-Mobile. O autor realça a importância de a realização das marcas manterem uma presença ativa nas redes sociais, pois estas são uma fonte valiosa de informação sobre os clientes. (Tan et al., 2020)

3. Enquadramento Metodológico

A abordagem de investigação usada na presente dissertação enquadra-se num estudo de caso. O estudo de caso é um método que tem como objetivo compreender fenómenos complexos. Podemos considerar um estudo de caso como " uma investigação empírica que estuda um fenómeno contemporâneo dentro do contexto de vida real de vida, especialmente quando as fronteiras entre o fenómeno e o contexto não são absolutamente evidentes"(Yin, 2011).

Os estudos de caso são utilizados quando o fenómeno é amplo e complexo, o conhecimento existente é insuficiente para se colocar questões causais , é necessária uma investigação aprofundada e o fenómeno não pode ser estruturado fora do contexto.

Como referido em capítulos anteriores o *Text Mining* tem como principal objetivo:

- Classificação de textos.
- Organização de grandes volumes de dados.
- Extração e recuperação de informação.
- Obtenção de novos conhecimentos/informação útil (Turban, 2015).

3.1 Análise de Sentimentos Caraterização

3.1.1 Arquitetura

A figura seguinte apresenta a arquitetura, seguida nesta dissertação.

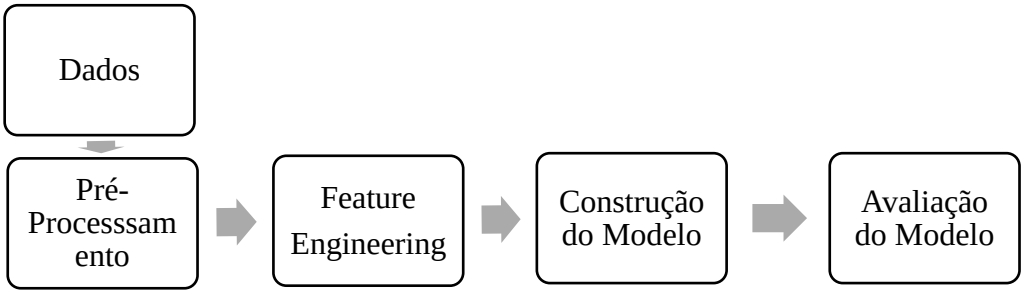


Figura 5- Arquitetura

O processo inicia-se com a recolha de um *dataset* que contém *reviews*[10] acerca de produtos, em seguida iremos proceder ao pré-processamento do texto. No pré-processamento do texto são aplicadas um conjunto de técnicas de pré-processamento em pipeline (tokenização, remoção de *stopwords*, *stemming*, etc.). Posteriormente é necessário representar o texto numa forma vetorial passível de ser entendida pelo classificador a que chamamos *Feature Engineering*. De seguida procedemos à construção do modelo (foram construídos diversos modelos usando diversos classificadores). Por último foi efetuada a avaliação do modelo construído através de um conjunto de métricas de avaliação do desempenho do modelo gerado.

As avaliações, comentários, *tweets* feitos pelos consumidores são, nos dias que correm, cada vez mais frequentes e úteis para quem procura um produto ou serviço. Permitem transmitir uma ideia de qualidade ou utilidade dos mesmos, a confiança no produto e no vendedor é fundamental no momento da compra online.

Em 2022 as estatísticas apontam que são processados 500 milhões de tweets postados todos os dias (Ahlgren, 2022). Devido a estes números elevados de utilizadores é de extrema importância fazer análise de sentimentos a partir de *tweets*, para deste modo, determinar o que é dito pelos utilizadores/compradores em relação a uma determinada marca ou serviço.

3.1.2 Pré-Processamento

A primeira etapa passou pela identificação e recolha de um *dataset*[11]. Os dados utilizados foram obtidos através da plataforma *Kaggle*. O objetivo será analisar dados não estruturados. O *dataset* utilizado contém comentários sobre produtos da *Amazon*.

Ao remover atributos sem grande relevância na classificação de sentimentos, ou seja, que não expresse uma opinião sobre o produto ou serviço, os problemas são minimizados e é otimizada a eficiência computacional.

A etapa de pré-processamento é de extrema importância pois irá eliminar palavras que não agregam muito valor ou informação (Benevenuto, Fabrício et al., 2015). Deste modo após a recolha de dados, procedeu-se ao pré-processamento de texto com a utilização de técnicas tais como:

[10] Análise critica em relação a algo
[11] Base de dados especificas

3.1.2.1 Conversão para minúsculas

Transforma todas as maiúsculas em minúsculas, pois as palavras minúsculas e maiúsculas são tratadas de forma diferente pela máquina com o objetivo de uniformizar a escrita.

3.1.2.2 Tratar contrações

Tratar contrações significa eliminar as abreviaturas presentes nas *reviews*. Com por exemplo: *"you're*" aqui temos de transformar para "*you are*"

3.1.2.3 Remoção de pontuação

Remover os caracteres especiais como pontuação ou outros caracteres que não são letras, como podemos identificar na seguinte imagem:

Figura 6- Caracteres

3.1.2.4 *Tokenize* ou Atomização

Consiste em transformar o texto em *tokens*, ou seja, partir o texto em palavras/termos individuais, utilizando o espaço como delimitador. A Tokenização é o processo de partir um texto nos seus componentes mais elementares.

3.1.2.5 Remover *Stopwords*

São palavras não discriminantes, isto é, que em nada contribuem para a compreensão do conteúdo. Servem apenas para unir palavras no texto e são por isso usadas com bastante frequência, mas, não acrescentam nada a um processo de classificação. Como exemplo de *Stopwords* temos determinantes, proposições, adjetivos. Por exemplo: *the, in, then, that, and, or*,entre outros.

3.1.2.6 Stemming vs Lemmatization[12]

Neta fase, testamos os cenários possíveis a utilização da técnica de processamento *Stemming* e a *Lemmatization*.

[12] Determinação da forma canônica que irá servir como lema

Stemming consiste no processo de reduzir a palavra ao seu radical, ou seja, caso existam sufixos ou prefixos estes são removidos. O Algoritmo de Porter é o mais usado na língua inglesa.

A *Lemmatization* consiste no processo de retornar o lema da palavra, que representa a forma básica da palavra

No seguinte quadro, podemos verificar as diferenças entre as duas técnicas de pré-processamento.

Stemming		***Lemmatization***	
computer *compute* *computation*	*Comput*	*saw* *see* *seen*	*See*

Tabela 3 - Diferença entre Stemming e Lemmatization

3.1.2.7 Feature Enginering

No *feature engineering* o texto é representado como um conjunto de todas as palavras (*bag-of-words*). Constrói-se um *vector of features* que se assemelha ao texto. Nesta fase o *bag-of-words* é convertido em números com algoritmos de aprendizagem máquina.

Neste estudo de caso utilizaremos dois vetores:

- *CountVectorizer*: Este método transforma um documento em vetores, em que conta as ocorrências de cada palavra em cada documento (Assery et al., 2019)
- *TfidfVectorizer*: (*Term Frequency-Inverse Document Frequency*) Utiliza a ponderação. Um dos problemas nas contagens simples é que, algumas palavras consideradas irrelevantes para a análise de sentimentos são contadas muitas vezes. O *TfidfVectorize* calcula a relevância das palavras nos dados. Assim, atribui maior peso para termos mais importantes e menor peso para os sem importância (Vel S., 2021).

3.1.3 Classificadores

A criação do modelo tem como propósito a classificação de sentimentos, ou seja, tem o objetivo de classificar os documentos como positivos ou negativos. Os seguintes classificadores que utilizamos são treinados e testados para obter a melhor previsão.

Deste modo, o principal objetivo é produzir um modelo que permita mapear todos os comentários em duas categorias (positivas ou negativas). A abordagem é baseada em aprendizagem supervisionada.

Quanto aos classificadores utilizados, seguem uma implantação e as mesmas métricas de avaliação. Começamos por invocar o modelo e posteriormente o cálculo da taxa de exatidão. Em seguida, procede-se à criação de uma matriz de confusão onde são obtidos os valores da precisão, sensibilidade, taxa de erro, *f-measure* e o *coeficiente kappa.* De referir que, quando o valor do *coeficiente kappa* é inferior a 60% não devemos considerar o valor da taxa de exatidão.

Após a revisão de literatura, aplicamos 7 classificadores que eram mais utilizados na análise de sentimentos, com o objetivo de obter a melhor taxa de exatidão, só ao testarmos vários modelos é que podemos afirmar qual o que melhor se comporta. Deste modo os classificadores que utilizamos foram, *SGD, KN Neighbors, Decision Tree, Linear SVM,* Bernoulli (Gaussian Naïve Bayes)*, Multinomial NB e por fim Logistic Regression.*

3.1.3.1 SGD (*Stochastic Gradient Descent*)

O SGD é um algoritmo que minimiza funções. O gradiente descendente começa com um conjunto inicial de valores de parâmetros e move-se em direção a um conjunto de valores de parâmetros que encontram o ponto mínimo para a função (Prasetijo et al., 2017).

É uma abordagem simples, mas, muito eficiente. (Cotter et al., 2022).

3.1.3.2 KN Neighbors (KNN)

Este algoritmo de classificação é baseado na proximidade dos vizinhos, mais concretamente na proximidade do vizinho mais próximo *(Nearest Neighbor*)(Bullejos et al., 2022).

O classificador KNN é um algoritmo de aprendizagem supervisionada.

Esta técnica tem como objetivo o reconhecimento de padrões através dos vizinhos que se encontrem mais próximos. O algoritmo KNN faz a classificação de uma dada instância, calculando a distância entre essa instância e os dados de treino. Isto é conseguido através

de uma função de cálculo de distância, que neste caso, é a distância Euclidiana.(H. Rodrigues, 2016)

O funcionamento de esta função é demonstrado na seguinte figura seguida:

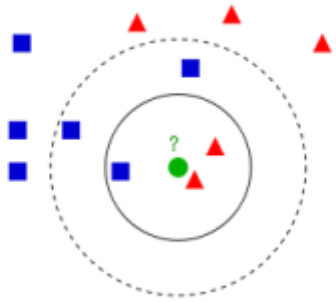


Figura 7- KN Neighbors Fonte: (H. Rodrigues, 2016)

Ao observar a figura, podemos concluir que o vizinho mais próximo seria os triângulos vermelhos.

3.1.3.3 Árvore de Decisão

O algoritmo Árvore de Decisão *(Decision Tree)*, o principal objetivo passo pela construção de uma árvore de testes que diferenciam documentos de classes diferentes, ou seja, uma decisão de classificação é uma sequência de testes em que estes resulta de uma classificação (H. Rodrigues, 2016).

A utilização da árvore de decisão para a classificação de uma frase ou documento é efetuada através da estrutura, desde a raiz até atingir uma determinada folha que representa o objetivo de classificar o documento

3.1.3.4 Linear SVM (*Support Vector Machine*)

O Support Vector Machine (SVM), representa um conjunto de técnicas e métodos de aprendizagem supervisionada usado para fins de classificação e regressão. Este classificador combina métodos estatísticos com métodos de aprendizagem máquina para gerar funções de mapeamento de input-output. Tem como input um vetor e como output zero ou um positivo ou negativo (Correia, Daniel, 2022).

O SVM é um algoritmo de classificação supervisionado que funciona bem para classificação de texto. A ideia principal do algoritmo SVM é construir um hiperplano que possa dividir em dois conjuntos. O hiperplano é como uma estrada que separa duas

categorias e leva em consideração a distância de recurso mais próxima do hiperplano. Este algoritmo representa as instâncias como pontos no espaço, mapeados para que os exemplos das diferentes categorias sejam separados por uma margem tão ampla quanto possível (Prasetijo et al., 2017).

A classificação é feita encontrando um cenário que diferencia as duas classes de forma eficiente, ou seja, que esta consiga delimitar as fronteiras das duas classes envolvidas com maior distância, como podemos verificar na figura seguinte.

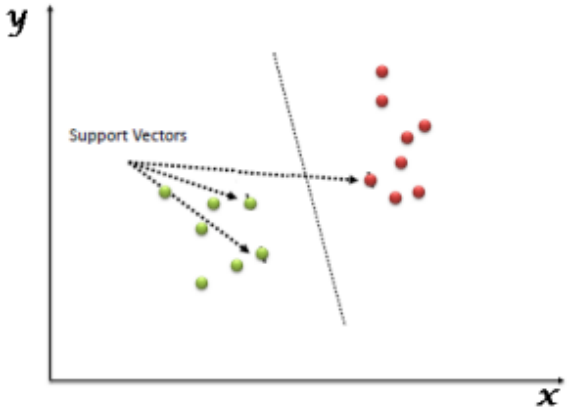


Figura 8- Support Vector Fonte: (H. Rodrigues, 2016)

3.1.3.5 Multinomial NB

O *Multinomial Naive Bayes* e um método de aprendizagem probabilístico. Na classificação de documentos, o objetivo é encontrar a melhor classe para o documento, neste casso para os comentários em relação a um produto (Silva, 2020).

O modelo Multinomial NB é projetado para determinar termo de frequência, ou seja, o número de vezes que um termo ocorre no documento. A frequência do termo também é útil para especificar se o mesmo é determinante na nossa análise. (Singh et al., 2019)

Por outro lado, às vezes, um termo pode ter uma grande presença num documento, ou seja repetir-se muitas vezes e pode ser uma palavra completamente irrelevante para análise em questão (Singh et al., 2019).

3.1.3.6 Bernoulli NB

Este algoritmo de aprendizagem máquina supervisionado. Assim como o *Multinomial NB,* este classificador é adequado para dados discretos.

No classificador Bernoulli NB, os recursos são variáveis binárias independentes que representa que se um termo está presente no documento em consideração ou não. Nesta abordagem, e ao contrário do classificador anterior, o objetivo é em apenas descobrir se um termo está presente ou ausente do documento em análise (Singh et al., 2019).

3.1.3.7 Regressão Logística

O algoritmo de regressão logística (LR) é também um dos algoritmos mais usados e mais avançados dentro da aprendizagem máquina, para classificação binária, ou seja, classificação positiva ou negativa. Este algoritmo mede a relação entre a variável dependente (aquela que pretendemos prever) e as restantes variáveis dependentes, estimando as probabilidades usando a sua função logística. Este modelo está relacionado com o modelo de regressão linear, mas usa uma função logística inversa para transformar o valor do output num valor entre zero e um, valor este que pode ser interpretado como uma variável.(M. Rodrigues, 2019).

Este modelo também apresenta desvantagem, diversos casos não podem ser previstos corretamente, visto que, a relação entre as variáveis não é sempre linear.

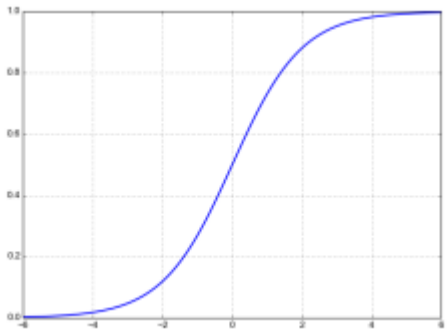


Figura 9- Regressão Logística Fonte:(M. Rodrigues, 2019)

3.1.4 Avaliação do Modelo

Esta funcionalidade avalia os resultados da classificação feita anteriormente. Através do cálculo da precisão *(precision)*, cobertura *(recall)*, taxa de acerto (*classification accuracy*) e da medida F *(F Measure*). Estes resultados permitem-nos avaliar se os algoritmos de classificação usados tiveram um bom desempenho ou não.

3.1.4.1 Taxa de Exatidão

A taxa de exatidão (*accuracy),* permite calcular a proporção de resultados corretamente classificados e é calculada da seguinte forma:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

1

Sendo relevante compreender os seguintes conceitos:

TP= Verdadeiro Positivo. Corresponde ao número de documentos que foram corretamente classificados.

TN = Verdadeiro Negativo. Corresponde ao número de documentos que foram corretamente classificados como não pertencendo à classe

FP = Falso Positivo. Corresponde ao número de documentos que foram incorretamente classificados como pertencendo à classe.

FN = Falso Negativo. Corresponde ao número de documentos que foram incorretamente classificados como não pertences da classe, quando na realidade pertenciam.

3.1.4.2 Sensitivity

A Sensibilidade (*sensitivity),* permite avaliar a se o classificador consegue identificar exemplos positivos, onde este é calculado da seguinte forma:

$$sensitivity = \frac{TP}{TP + FN}$$

2

3.1.4.3 Specificity

A *specificity*, ao contrário da *sensitivity*, pretende avaliar a capacidade de o classificador identificar exemplos negativos, quanto maior esta for menor será a probabilidade de um exemplo ser classificado como positivo sendo este negativo e assim a taxa de falsos negativos será baixa. Esta é determinada da seguinte forma:

$$specificity = \frac{TN}{TN + FP}$$

3

3.1.4.4 Precision

A *precision* consegue avaliar a precisão de um classificador na identificação de exemplos positivos.

$$precision = \frac{TP}{TP + TN}$$

4

3.1.4.5 F-measure

A F-measure, é a métrica que é utilizada na área da extração de informação, permitindo uma média ponderada entre a *precision* e o *rcall*.(Nguyen et al., 2018).

A métrica é calculada da seguinte forma:

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall}$$

5

4. Caso de Estudo – *Amazon*

As avaliações, comentários, *tweets* feitos pelos consumidores são, nos dias que correm, cada vez mais frequentes e úteis para quem procura um produto ou serviço. Permitem transmitir uma ideia de qualidade ou utilidade dos mesmos, a confiança no produto e no vendedor é fundamental no momento da compra online.

Em 2022 as estatísticas apontam que são processados 500 milhões de tweets postados todos os dias (Ahlgren, 2022). Devido a estes números elevados de utilizadores é de extrema importância fazer análise de sentimentos a partir de *tweets*, para deste modo determinar o que é dito pelos utilizadores/compradores em relação a uma determinada marca ou serviço

A *Amazon* é uma das maiores empresas de comércio eletrónico do mundo. A *Amazon* oferece atualmente mais de 12 milhões de produtos diferentes. No site da *Amazon* podemos encontrar variadíssimos produtos tais como: livros, música, jogos, filmes, roupas, produtos eletrónicos, brinquedos entre outros.

A *Amazon* desde a sua criação até aos dias de hoje tem crescido e superado qualquer precisão. Atualmente, num período difícil com inflação elevada, os resultados da Amazon foram considerados bastante bons. Existe então um enorme interesse na análise de sentimentos das *reviews* de produtos (Guerra, 2022). Os *feedbacks* dos clientes têm elevada importância para as empresas, visto que, conseguem reconhecer os seus pontos fracos, pontos fortes e o que podem fazer para melhorar o produto ou serviço (Liu, 2012). Se por um lado é benéfico para uma organização saber do feedback dos clientes também importante para futuros clientes, para reduzir a incerteza no momento da decisão de efetuar a compra.

Em geral, a informação online partilhada pelos clientes tem duas formas típicas a classificação das críticas. Uma das formas é universal e a Amazon tem disponível no site. Esta forma permite que os clientes classifiquem um produto de 1 a 5 estrelas (em que o 1 é a avaliação mais baixa o 5 a melhor), e fornece um resumo da experiências e opiniões sobre um determinado produto. Por outro lado, tempos as críticas que são

comentários em forma de texto que consegue descrever a experiência do cliente mais detalhada e mais subjetivas.

Com o aparecimento de técnicas de *Text Mining*, conseguimos extrair a informação a partir de textos não estruturados de modo a agregar da melhor forma para conseguir classificar se o sentimento é positivo ou negativo.

O presente caso de estudo é baseado num *dataset* do site *Kaggel e* contém informação sobre *reviews* sobre produtos da *Amazon.* Cada opinião/ *review* presente no *dataset* já está classificada quanto à polaridade. Como o objetivo é a identificação das opiniões, sentimentos e emoções expressas num documento, deste modo as opiniões são classificadas quanto à polaridade, podem ser positivas, negativas e neutras

A escolha linguagem de programação acabou de recair sobre a ferramenta *Python* e o ambiente onde este foi desenvolvido foi o *Visual Studio Code* Uma das razões pelas quais foi feita esta escolha foi o facto de ser das mais usadas em contexto de aprendizagem máquina e mineração de texto. Possui, para além disso, um conjunto de bibliotecas, o que representa uma grande vantagem quando pretendemos construir um protótipo rapidamente. Estas bibliotecas permitem o desenvolvimento de código nas mais diversas áreas, sendo uma delas a Ciência dos Dados (*Data Science)*.

No nosso estudo, foi utlizado a biblioteca NLTK, é parte da linguagem em *Paython*, integrando um conjunto de algoritmos

Todas estas características da linguagem fazem com que seja uma linguagem que facilita muito na fase inicial da preparação do texto, ou seja, do pré-processamento textual.

4.1 Caracterização

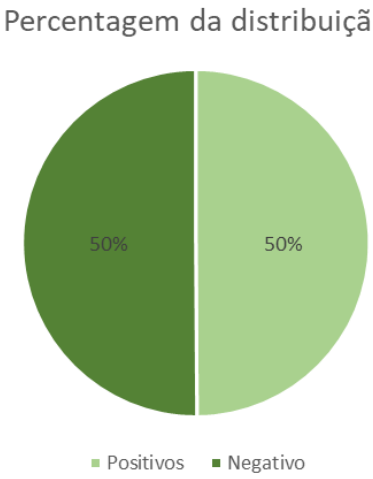
4.1.1 Caracterização do Conjunto de dados (Dataset)

O *dataset* analisado é composto por duas colunas, uma contém o texto de opinião [*Text*] e a outra a classificação [*Sentiment*]. O conjunto de dados contém um total de 1000 linhas, ou seja, é constituído por um total de 1000 avaliações.

	Text	Sentiment
0	So there is no way for me to plug it in here i...	negative
1	Good case Excellent value.	positive
2	Great for the jawbone.	positive
3	Tied to charger for conversations lasting more...	negative
4	The mic is great.	positive
5	I have to jiggle the plug to get it to line up...	negative
6	If you have several dozen or several hundred c...	negative
7	If you are Razr owner...you must have this!	positive
8	Needless to say I wasted my money.	negative
9	What a waste of money and time!.	negative

Figura 10 - Classificação dataset

Em termos de sentimentos positivos e negativos, estão bastante balanceados como podemos verificar nas imagens seguintes:




```
---------Número de reviews positivas e negativas------
negative    492
positive    490
```

Figura 11 - Número de reviews positivas e negativas

Figura 12- Percentagem da distribuição por tipo de review

Num total de 1000 avaliações, sendo 492 negativas e 490 positivas, sobram ainda 18 valores que estavam classificados com o sentimento neutro. O que resulta em cerca de 50% são classificados como positivos e 50% como negativos podemos afirmar que o *dataset* é balanceado. Deste modo, consideramos que não é necessário aplicar técnicas de balanceamento.

É importante referir que antes de realizar qualquer tratamento de dados, tínhamos na nossa base de dados um total de 10244 palavras.

No início desta metodologia e para nos ajudar a conhecer melhor o *dataset* foi elaborado um gráfico *(Wordcloud)* de palavras onde é possível, visualmente, representar as palavras que podemos encontrar no *dataset* analisado.

Conseguimos verificar, a existência de elevados elementos que não nos permitem determinar se um sentimento é positivo ou negativo, algo que é normal visto que não procedemos a nenhum processamento de texto.

Figura 13- Nuvem de Palavras com o dataset original

Na fase inicial, foi elaborado um gráfico com a frequência de palavras, os 10 termos mais frequentes presente na base de dados. Ao observar o gráfico abaixo apresentado, podemos concluir que os termos mais frequentes são termos irrelevantes para a análise de sentimentos. Com estes termos, não conseguimos determinar se um sentimento é negativo ou positivo, por isso é necessário proceder ao pré-processamento textual, para remover palavras irrelevantes que em nada contribuem para a classificação de sentimentos.

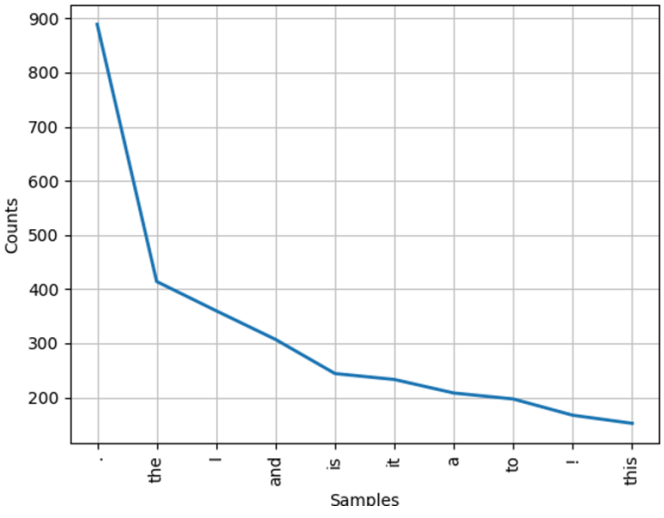


Gráfico 1- Gráfico de Frequências de Palavras do dataset original

4.1.2 Pré-Processamento

Aplicamos diversas técnicas de pré-processamento de dados. Esta etapa é fundamental para podermos prosseguir para a criação do modelo de classificação do sentimento (positivo, negativo) que enunciamos de seguida.

4.1.2.1 Conversão para minúsculas

O primeiro passo efetuado foi transformar o texto em letras minúsculas para uniformizar o texto. Com a aplicação desta técnica de pré-processamento não iremos obter qualquer impacto no número de palavras, visto que só vai converter maiúsculas em minúsculas.

```
review_df["Text"] = review_df["Text"].str.lower()
print(review_df.head(10))
```

Figura 14 - Conversão para minúsculas

4.1.2.2 Tratamento das contrações

No tratamento das contrações, eliminamos abreviaturas presentes no *dataset d*ado que trabalhamos no domínio do inglês. Como por exemplo *"you're"* passa para *"you are"*. Ao aplicar esta técnica de pré-processamento verificamos um aumento do número de termos.

```
review_df['Text'] = review_df['Text'].apply(lambda x: contractions.fix(x))
print(review_df.head(10))
```

Figura 15 - Tratar contrações

4.1.2.3 Remoção da pontuação e caracteres especiais

Remoção de todos os caracteres especiais como a pontuação, esta função vai substituir a pontuação por um caracter vazio. O objetivo desta técnica visa excluir todos os caracteres especiais e a pontuação, visto que, neste estudo de caso são irrelevantes para a classificação do sentimento. Um ponto final ou uma virgula podem ser informativos, mas, neste estudo de caso iremos removê-los.

```
review_df['Text'] = review_df['Text'].str.replace(r'[^\w\s]+', '')
print(review_df.head(10))
```

Figura 16 - Remoção Pontuação

4.1.2.4 Tokenize

Nesta fase, procedemos à *tokenização,* o objetivo é a separação do texto nos seus componentes mais elementares que são as palavras utilizando o espaço como caracter delimitador para separar as mesmas.

```
#review_df['tokenized_text'] = ' '.join( review_df['text'].apply(word_tokenize))
review_df['tokenized_text']  = df.apply(lambda row: nltk.word_tokenize(row['Text']), axis = 1)
```

Figura 17- Tokenize

4.1.2.5 Remover *Stopwords*

A remoção de *Stopwords* também foi aplicada no sentido de eliminar tudo o que não nos permite determinar se o sentimento é positivo ou negativo.

Consiste na eliminação de verbos auxiliares, artigos, pronomes, termos mais comuns e irrelevantes para a classificação.

```
review_df['Text'] = review_df['Text'].apply(lambda x: ' '.join([word for word in x.split() if word not in (stop)]))
print(review_df.head(10))
```

Figura 18 - Remover Stopwords

4.1.2.6 Stemming vs Lemmatization

Nesta fase decidimos construir dois cenários possíveis e testar qual o que melhor se comporta, com este tipo de dados, para desta forma avaliar o que produz melhores resultados.

O processo de *stemming* , as palavras são reduzidas ao ser radical, com vista a não serem tratadas como palavras diferentes. Esta técnica de pré-processamento elimina os prefixos e sufixos de cada termo (palavra).

O processo de lematização permite-nos obter a raiz morfológica da palavra, ou seja, o lema da palavra.

Ao testarmos individualmente cada técnica de pré-processamento iremos obter resultados diferentes, por este motivo iremos testas as duas técnicas e analisar qual a que produz mais efeito no nosso *dataset*.

Aplicamos isoladamente cada uma das técnicas. Começando pela aplicação de *Stemming*, pretendemos reduzir as palavras ao seu radical, para garantir que as palavras originais e as suas derivações sejam tratadas da mesma forma.

Na técnica de pré-processamento *Lemmatization*, o objetivo é retirar as inflexões e obter a forma base das palavras.

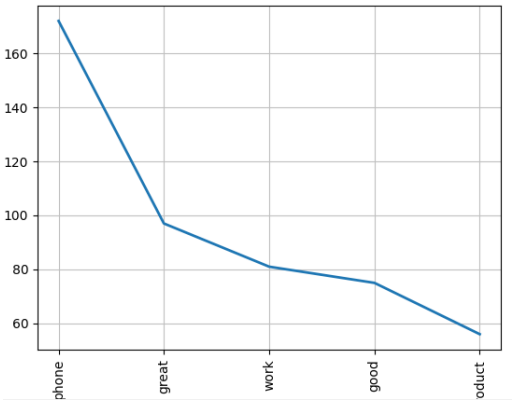
```
#stemming porter stemmer
porter = PorterStemmer()
review_df['tokenized_text'] = review_df['tokenized_text'].apply(lambda x: [porter.stem(y) for y in x])
print(review_df.head(10))
#------ Número de Palavras
print (len(' '.join(map(str,review_df["tokenized_text"])).split()))
```

Figura 19 - Aplicação de Stemming

```
#Lemmatização
from nltk.stem.wordnet import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
review_df['tokenized_text'] = review_df['tokenized_text'].apply(lambda x: [lemmatizer.lemmatize(y) for y in x])
print("Lemmatização:",review_df.head(10))
#------ Número de Palavras
print (len(' '.join(map(str,review_df["tokenized_text"])).split()))
```

Figura 20 - Aplicação de Lemmatização

Ao comparamos os diferentes *outputs*, verificamos que o número de palavras após a aplicação das duas técnicas de pré-processamento foi idêntico. Todavia, quando comparamos a frequências das palavras aí obtemos algumas diferenças



"phone": 172

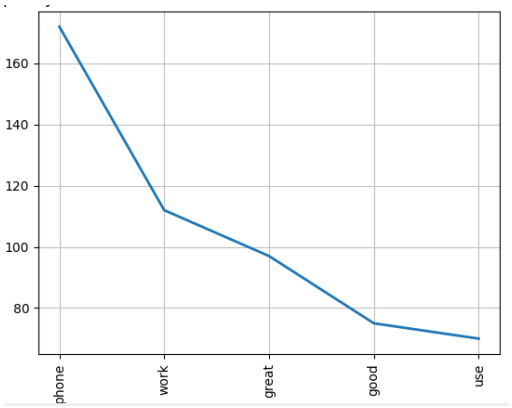
"work": 81

"great": 97

"good":75

"product": 56

Gráfico 2- Gráfico de Frequência aplicando Lematização



"phone": 172

"work": 112

"great": 97

"good":75

"use": 70

Gráfico 3 - Gráfico de Frequência aplicando Setemming

Deste modo, a diferença encontrada não nos permite dizer com certeza se um método de pré processamento é mais eficaz do que o outro a nível da redução do número de palavras. Para podermos dizer qual é a melhor técnica de pré-processamento o *stemming* ou a lematização temos de recorrer ao classificador e analisar qual a que tem melhor taxa de exatidão e aí identificar qual a melhor técnica.

Na seguinte tabela apresentamos o quadro com todas a técnicas e etapas de pré-processamento de texto utilizados.

Fases	Nº de Tokens	%	Nº Tokens
Inicial	11696	0%	0
Tokenização	10244	-15%	-1452
Conversão p/Minúsculas	10244	0%	0
Tratar Contrações	10420	2%	176
Remoção Pontuação	10373	0%	-47
Remoção de Stopwords	5254	**-51%**	**-5119**
Stemming vs Lemmatização	5242	0%	-12

Tabela 4 - Tabela resumo com aplicação das diversas técnicas do Pré-processamento

Observando a tabela anterior conseguimos concluir, que o método de pré-processamento de texto que produziu mais efeito na nossa base de dados foi, claramente, a remoção de *stopwords*, esta técnica de pré -processamento de texto reduziu em mais de metade o número de palavras (-51%), passando de 10373 *tokens* para 5254.

Depois de aplicadas todas as técnicas de pré-processamento voltamos a elaborar uma nuvem de palavras.

Figura 21 - Nuvem de Palavras no final do pre-processamento

Ao analisar a figura anterior (nuvem de palavras) podemos constatar que a palavras que se encontram com maior destaque, são palavras que nos permite determinar se o sentimento presente nas *reviews* é positivo ou negativo.

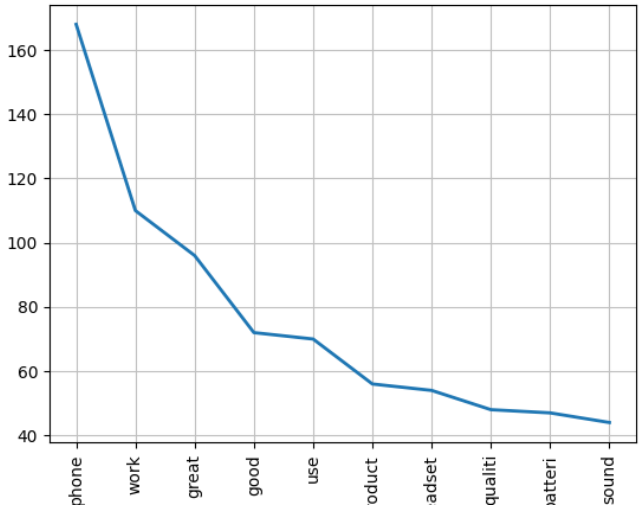


Gráfico 4 - Os 10 Termos mais frequentes após pre-processamento

Com a análise do gráfico é evidente uma grande alteração face ao primeiro gráfico de palavras apresentado antes de iniciar as etapas de pré-processamento.

Este gráfico, espelha a frequência que obtivemos na nuvem de palavras anterior, identificamos palavras com maior tamanho, ou seja, com maior frequência *"phone", "work", "great", "good", "use", "product"* entre outras.

4.1.3 Classificadores

Para que treinar e testar os classificadores é necessário, em primeiro lugar dividir o *dataset* em dados de treino e dados de teste.

70%: Foi atribuído para treino

30%: Foi atribuído para teste

Os algoritmos de classificação de sentimentos foram usados para classificar o documento como positivos ou negativos, o objetivo passa por avaliar a capacidade de erro e acerto dos nossos modelos, mas antes de usarmos os classificadores precisamos de transformar os nossos dados numa matriz através do vetorizado e só depois utilizar o modelo treinado

O primeiro apenas conta o número de vezes que uma palavra aparece, o que faz com que favorece, as palavras que aparecem muitos vezes e o segundo considera quantidade total de palavras no cálculo da frequência de palavras. O *TfidfVectorizer()* é conhecido por ser o mais popular, pois, utiliza a frequência das palavras.

O *TfidfVectorizer* o peso de um termo calcula-se multiplicando a frequência do termo no documento pela sua probabilidade inversa de o termo ocorrer no conjunto de documentos. Um termo que ocorra com mais frequência no documento e outro que ocorra com menos frequência na coleção de documentos é determinado com menor peso e pouco poder de discriminação. A frequência inversa do documento de uma palavra é medida da importância relativa em todo o conjunto (*corpus)*(Nguyen et al., 2018).

$$TF(palavra) = \frac{Frequência\ da\ palavra\ no\ documento}{Número\ de\ palavras\ no\ documento}$$

6

$$IDF(palavra) = \log\left(\frac{Número\ total\ de\ documentos}{Número\ de\ docuemntos\ que\ contém\ a\ palavrs}\right)$$

7

O *TfidfVectorizer* é o produto da equação 6 com a equação 7, cada documento é representado como um vetor que contém as pontuações para cada um das palavras no documento.(Nguyen et al., 2018) Este utiliza a frequência do termo no documento pela ponderação, o que significa que se uma palavra aparece em muitos textos, provavelmente,

é uma palavra comum e não relevante na avaliação do texto (Dastani et al., 2020). Em *Python, a função TfidfVectorizer* converte o documento em uma matriz de atributos

Também utilizamos o *CountVectorizer()* para transformar os dados de texto numa matriz de valores. Este apenas conta o número de vezes que uma palavra (termo) aparece na frase

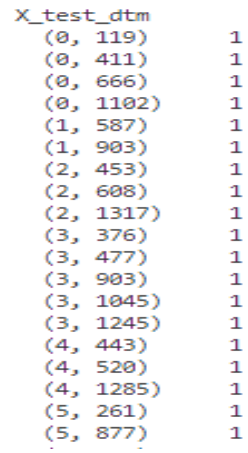


Figura 22- Exemplo da Matriz

Inicialmente, aplicamos todos os classificadores utilizando a técnica de pré-processamento *stemming* e depois voltamos a aplicar os mesmos classificadores, mas desta vez com a técnica de pré-processamento *Lematização,* para perceber se esta tem algum positivo na avaliação de métricas.

O primeiro classificador utilizado foi a Regressão Logística (RL), tendo sido usado o método *Logistic Regression ()* do *sklearn.linear_model*.

Utilizamos também o classificador de arvore de decisão *(Decision Tree),* com a utilização dos modelos *DecisionTreeClassifier ()* do *sklearn.tree.*

Para as *Support Vector Machines (SVM)* usamos o modelo *SVC () do sklearn.SVM*.

No caso do modelo *K-Nearest Neighbours (KNN)* utilizamos o classificador *KNeigborsClassifier ()* do *sklearn.neighbors*.

Aplicamos também o modelo de classificação *Stochastic Gradient Descent (SGD)* invocando *sklearn.linear_model.SGDClassifier*

Para o *Bernoulli (NB)* usamos o modelo *sklearn.navie_bayes*

E por fim o Multinomial (NB), o classificador *Naive bayes* para modelos multinomiais usando *o sklearn.navie_bayes.MutinomialNB.*

4.1.3.1 Regressão Logística

Como referido, anteriormente, começamos por aplicar o classificador de regressão logística com a técnica de pré-processamento *stemming.*

Nas seguintes tabelas, podemos observar as métricas de avaliação dos classificadores e a matriz de confusão respetivamente.

Tx.Exatidão	Precisão	Sensibilidade	*F-measure*	Tx. Erro	*Kappa*
0.83	0.85	0.81	0.81	0.18	1

Tabela 5- Métrica Avaliação RL

		Detetada	
		Sim	Não
Real	Sim	TP=116	FN=27
	Não	FP=25	TN=127

Tabela 6 - Matriz de Confusão Regressão Logística

4.1.3.2 *KN Neighbors*

Tx.Exatidão	Precisão	Sensibilidade	*F-measure*	Tx. Erro	*Kappa*
0.64	0.6	0.84	0.69	0.36	0.3

Tabela 7- Métricas de Avaliação KNN

		Detetada	
		Sim	Não
Real	Sim	TP=194	FN=35
	Não	FP=141	TN=121

Tabela 8 - Matriz de Confusão KN Neighbors

4.1.3.3 Árvore de Decisão

Tx.Exatidão	Precisão	Sensibilidade	*F-measure*	Tx. Erro	*Kappa*
0.79	0.78	0.79	0.78	0.2	0.53

Tabela 9 - Métricas de Avaliação DT

		Detetada	
		Sim	Não
Real	Sim	TP=112	FN=29
	Não	FP=31	TN=123

Tabela 10 - Matriz de Confusão Árvore de Decisão

4.1.3.4 Linear SVM

Tx.Exatidão	Precisão	Sensibilidade	*F-measure*	Tx. Erro	*Kappa*
0.78	0.77	0.80	0.78	0.21	0.59

Tabela 11 - Métricas de Avaliação SVM

		Detetada	
		Sim	Não
Real	Sim	TP=113	FN=28
	Não	FP=31	TN=120

Tabela 12 - Matriz de Confusão Linear SVM

4.1.3.5 Multinominal NB

Tx.Exatidão	Precisão	Sensibilidade	*F-measure*	Tx. Erro	*Kappa*
0.78	0.74	0.86	0.80	0.22	0.50

Tabela 13 - Métricas de Avaliação MNB

		Detetada	
		Sim	Não
Real	Sim	TP=125	FN=20
	Não	FP=44	TN=106

Tabela 14 - Matriz de Confusão Multinominal NB

4.1.3.6 Stochastic Gradient Descent SGD

Tx.Exatidão	Precisão	Sensibilidade	*F-measure*	Tx. Erro	*Kappa*
0.81	0.81	0.75	0.78	0.22	0.53

Tabela 15 - Métricas de Avaliação SGD

		Detetada	
		Sim	Não
Real	Sim	TP=119	FN=38
	Não	FP=31	TN=107

Tabela 16 - Matriz de Confusão SGD

4.1.3.7 *Bernoulli NB*

Tx.Exatidão	Precisão	Sensibilidade	*F-measure*	Tx. Erro	*Kappa*
0.8	0.77	0.86	0.81	0.19	0.61

Tabela 17 - Métricas de Avaliação Bernoulli NB

		Detetada	
		Sim	Não
Real	Sim	TP=126	FN=18
	Não	FP=43	TN=108

Tabela 18 - Matriz de Confusão Bernoulli NB

Na fase de pré-processamento de texto, foi notório que a utilização de a técnica de *stemmig* e de *Lematização* produziu o mesmo efeito na redução do número de palavras.

Deste modo, até aqui utilizamos a técnica de pré-processamento *stemming*, e agora iremos trocar essa mesma técnica para a lematização, para analisar se, nesta etapa, produz diferenças sejam elas positivas ou negativas.

4.1.3.8 Regressão logística c/*lematização*

Tx.Exatidão	Precisão	Sensibilidade	*F-measure*	Tx. Erro	*Kappa*
0.77	0.80	0.74	0.77	0.23	1

Tabela 19 - Regressão logística c/stemming

		Detetada	
		Sim	Não
Real	Sim	TP=113	FN=39
	Não	FP=29	TN=114

Tabela 20 - Matriz de Confusão c/lematização

4.1.3.9 KN Neighbors c/ *lematização*

Tx.Exatidão	Precisão	Sensibilidade	*F-measure*	Tx. Erro	*Kappa*
0.62	1	0.84	0.67	0.38	0.26

Tabela 21 - KN Neighbors c/stemming

		Detetada	
		Sim	Não
Real	Sim	TP=193	FN=36
	Não	FP=150	TN=112

Tabela 22 - Matriz de confusão KN Neighbors c/lematzação

4.1.3.10 Árvore de decisão c/ *lematização*

Tx.Exatidão	Precisão	Sensibilidade	*F-measure*	Tx. Erro	*Kappa*
0.76	0.78	0.69	0.73	0.25	0.50

Tabela 23 - Arvore de decisão c/stemming

		Detetada	
		Sim	Não
Real	Sim	TP=99	FN=45
	Não	FP=28	TN=123

Tabela 24 - Árvore de Decisão c/lematização

4.1.3.11 Linear *SVM* c/ *lematização*

Tx.Exatidão	Precisão	Sensibilidade	*F-measure*	Tx. Erro	*Kappa*
0.83	0.86	0.80	0.83	0.17	0.66

Tabela 25 - Linear SVM c/stemming

		Detetada	
		Sim	Não
Real	Sim	TP=121	FN=30
	Não	FP=20	TN=124

Tabela 26 - Matriz de confusão Linear SVM c/lematização

4.1.3.12 Multinomial NB c/ *lematização*

Tx.Exatidão	Precisão	Sensibilidade	*F-measure*	Tx. Erro	*Kappa*
0.78	0.71	0.89	0.79	0.22	0.56

Tabela 27 - Multinomial NB c/stemming

		Detetada	
		Sim	Não
Real	Sim	TP=122	FN=15
	Não	FP=51	TN=107

Tabela 28 - Matriz de Confusão Multinomial NB c/ lematização

4.1.3.13 SGD c/ lematização

Tx.Exatidão	Precisão	Sensibilidade	*F-measure*	Tx. Erro	*Kappa*
0.76	0.76	0.73	0.74	0.24	0.57

Tabela 29 - SGD c/stemming

		Detetada	
		Sim	Não
Real	Sim	TP=104	FN=39
	Não	FP=33	TN=119

Tabela 30 - Matriz de Confusão SGD c/lematização

4.1.3.14 Bernoulli NB c/ lematização

Tx.Exatidão	Precisão	Sensibilidade	*F-measure*	Tx. Erro	*Kappa*
0.78	0.71	0.89	0.79	0.22	0.56

Tabela 31 - Bernoulli NB c/stemming

		Detetada	
		Sim	Não
Real	Sim	TP=126	FN=26
	Não	FP=42	TN=101

Tabela 32 - Matriz de confusão Bernoulli NB c/ lematização

4.2 Comparação dos classificadores

Nesta fase, inicia-se a medição da qualidade dos algoritmos descritos, anteriormente, com o objetivo de avaliar qual o classificador que melhor se comporta com os nossos dados. Avaliamos a capacidade de erro e o acerto do nosso modelo. Como referimos anteriormente dividimos os nossos dados em dois subconjuntos isolados e independentes em que 70% dos dados foram atribuídos para treino e 30% para teste

De modo a confirmar o algoritmo que melhor se comporta com os nossos dados e que obtém uma previsão mais precisa. Calculámos diversas métricas de avaliação de modelos. Como referimos anteriormente, criamos dois cenários.

1ª Cenário

Neste primeiro cenário, na etapa de pré-processamento mantivemos a técnica de pré-processamento *stemming* e de seguida medimos a qualidade dos algoritmos e calculámos as diversas métricas descritas anteriormente.

Os resultados obtidos, das diferentes métricas de avaliação estão presentes nos seguintes gráficos.

Gráfico 5 - Taxa de Exatidão c/stemming

A taxa de exatidão, com a utilização da técnica de pré-processamento stemming, obteve melhores resultados com o classificador de Regressão logística 83%. O por classificador foi o *KNN* com 64%.

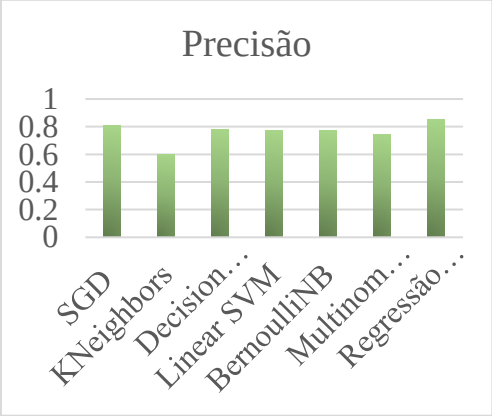


Gráfico 6 - Precisão c/Stemming

A precisão obteve resultados idênticos coma taxa de exatidão. A Regressão logística consegui ter o melhor resultado 85%. O KNN voltou a ser o pior com 60%.

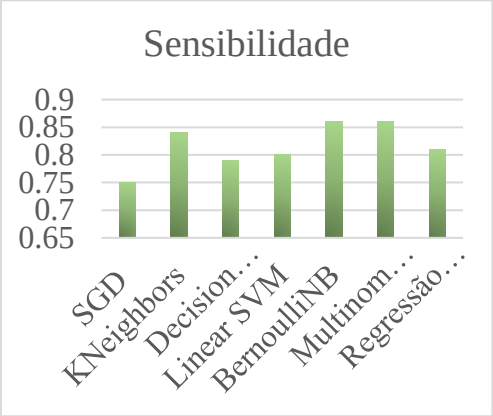


Gráfico 7 - Sensibilidade c/Stemming

Com a análise da sensibilidade, o panorama já é um pouco diferente, visto que, o que obteve melhor resultado foi o classificador Bernoulli NB, o SGD obteve o resultado mais baixo 75%.

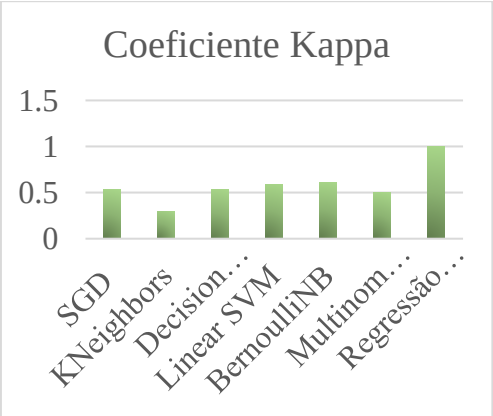


Gráfico 8 - Coeficiente Kappa c/Stemming

Para podermos considerar a taxa de exatidão como certa, o coeficiente *kappa*, que é uma medida de concordância, ou seja, mede o grau de concordância além do que seria esperado (Mincov et al., 2022). O coeficiente *kappa* tem de estar acima dos 60%, o que não acontece o *SGD*, o *KNN*, a árvore de decisão, *SVM* e o *Multinomial NB*. A regressão logística apresenta o valor mais alto de *kappa*

Gráfico 9 - Taxa de erro c/Stemming

Com a taxa de erro, podemos verificar que a Regressão logística e o BernoulliNB apresentam o mesmo resultado 19% de taxa de erro. O KNN tem a taxa de erro mais elevada 36%.

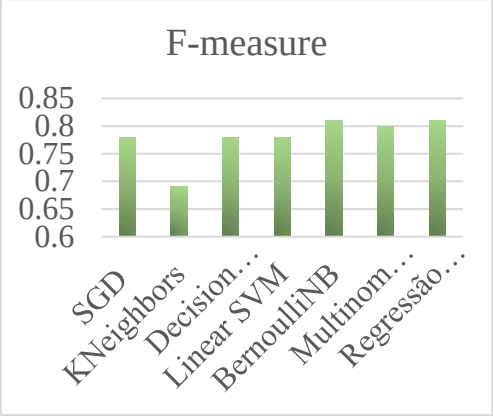


Gráfico 10 - F-measure c/Stemming

O *f-measure* apresenta valores mais elevados com a Regressão Logística e com o BernoulliNB 81%. O KNN volta a obter o pior resultado nesta métrica de avaliação 69%.

Graficamente, conseguimos inferir que o algoritmo que tem melhor resultado em termos de taxa de exatidão é a Regressão Logística com a técnica de pré-processamento *stemming.* O classificador Regressão logística é o que melhor se comporta tanto a nível de taxa de exatidão (0.83), precisão (0.85), *f-measure (0.81),* taxa de erro (0.18), e o *coeficiente kappa (1).*

Por outro lado, quando analisamos a sensibilidade, o classificador que obtém o melhor resultado é o *Multinomial NB* (0.86).

Após a observação destes dados, podemos também concluir que o classificado que obtém os piores resultados é o *K-Nearest Neighbor*.

2ª Cenário

No segundo cenário, na etapa de pré-processamento trocamos a técnica de pré-processamento *stemming* e colocamos *a lematização.* De seguida, voltamos a medir a qualidade dos algoritmos e calculámos as diversas métricas descritas anteriormente, para apurar se conseguimos obter melhores resultados com esta técnica de pré-processamento

Os resultados obtidos, das diferentes métricas de avaliação, estão presentes nos seguintes gráficos.

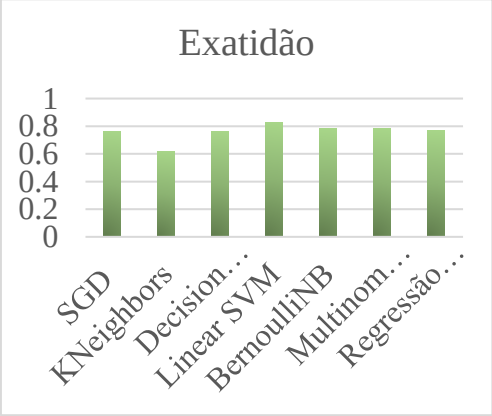


Gráfico 11 - Taxa de Exatidão c/Lematização

A taxa de exatidão com a aplicação da técnica de pré-processamento Lematização, apresenta um melhor resultado com o classificador SVM 83%.O KNN, tal como no primeiro cenário, apresenta o pior resultado com 62%.

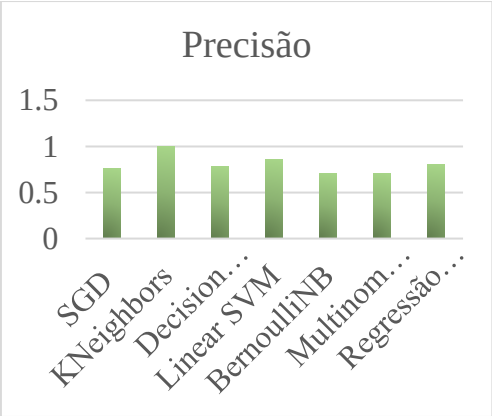


Gráfico 12 - Precisão c/Lematização

No que diz respeito à precisão, o KNN tem o valor mais alto e de seguida o SVM com 86%.

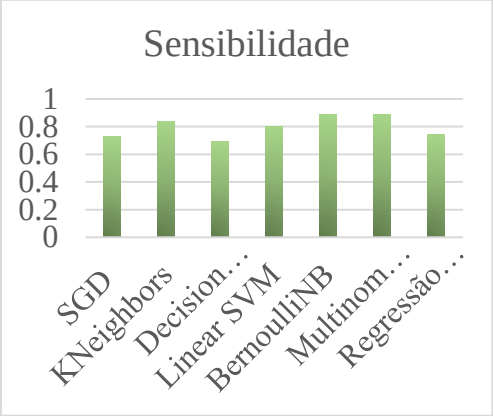


Gráfico 13 - Sensibilidade c/Lematização

O BernoulliNB e o MultinomialNB apresentam o mesmo valor na métrica de avaliação sensibilidade (89%). O valor mais baixo pertence à Àrvore de Decisão com 69%.

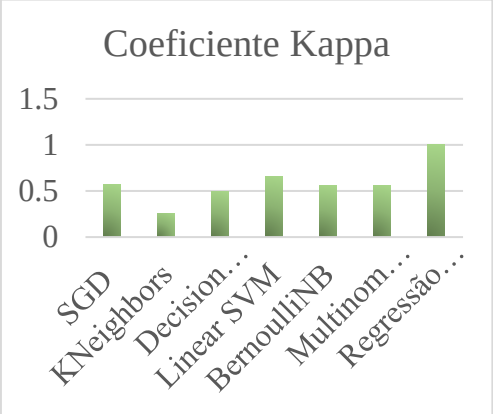


Gráfico 14 - Coeficiente Kappa

Como abordado anteriormente, se o *kappa* for inferior a 60%, deixamos de considerar o valor da taxa de exatidão. O KNN obteve a melhor taxa de exatidão, mas, por outro lado teve o pior coeficiente *kappa* 26%, ou seja, não podemos considerar a taxa de exatidão. A Regressão Logística alcançou o melhor *kappa*.

Gráfico 15 - Taxa de Erro c/Lematização

A taxa de erro mais baixa pertence ao classificador SVM 17%. O KNN tem a taxa de erro mais elevada38%.

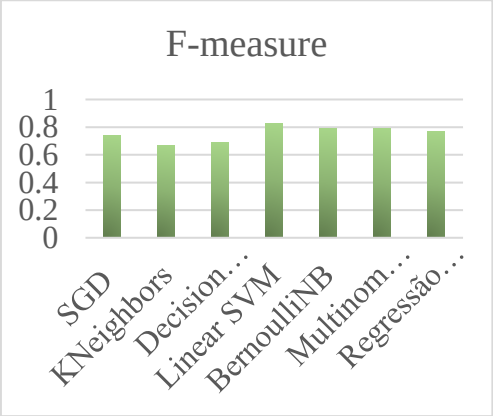


Gráfico 16 - F-measure c/Lematização

No que diz respeito ao *f-measure*, apresentam valores muito idênticos. O SVM alcançou o valor mais elevado comparando com os restantes de 83%.

Após a análise dos gráficos, acima representados, podemos inferir que existem diferenças na aplicação das duas técnicas de pré-processamento (*Stemming/Lematização*).

No primeiro cenário, como foi descrito anteriormente, o algoritmo que apresentava melhor resultados era o de Regressão Logística.

No segundo cenário apresentado, podemos verificar que tal não acontece.

Com a aplicação da técnica de pré-processamento Lematização constatamos que o algoritmo que obtém melhor resultado na maioria das métricas de avaliação é o *Support*

Vector Machines com uma taxa de exatidão de 83%, uma precisão de 86%, sensibilidade de 80%, o *f-measure* de 83%, o *coeficiente kappa* 0.66 e uma taxa de erro de 17%.

Um dado curioso, é que tanto no primeiro cenário como no segundo cenário o classificador *K-Nearest Neighbor*, no geral, é o que tem piores resultados. Com as maiores taxas de erro e com o *coeficiente kappa* mais alto.

Na seguinte tabela, podemos observar o resumo de todos os testes feitos nestes dois cenários:

	Regressão Logística		**SGD**		**DT**		**SVM**		**KNN**		**Multinomial NB**		**Bernoulli NB**	
Métricas de Avaliação	Stem	Lemt	Stem	Lemt	Stem	Lemt	Stem	Lemt	Stem	Lemt	Stem	Lemt	Stem	Lemt
Taxa de Exatidão	**0,83**	0,77	0,81	0,76	0,79	0,76	0,78	**0,83**	0,64	0,62	0,78	0,78	0,8	0,78
Coeficiente Kappa	**1**	1	0,53	0,57	0,53	0,5	0,59	**0,66**	0,3	0,26	0,5	0,56	0,61	0,56
Sensibilidade	**0,81**	0,74	0,75	0,73	0,79	0,69	0,8	**0,8**	0,84	0,84	0,86	0,89	0,86	0,89
F1-measure	**0,81**	0,77	0,78	0,74	0,78	0,73	0,78	**0,83**	0,69	0,67	0,8	0,79	0,81	0,79
Taxa de Erro	**0,18**	0,23	0,22	0,24	0,2	0,25	0,21	**0,17**	0,36	0,38	0,22	0,22	0,19	0,22
Precisão	**0,85**	0,8	0,81	0,76	0,78	0,78	0,77	**0,86**	0,6	1	0,74	0,71	0,77	0,71

Tabela 33- Tabela Resumo dos Classificadores

Após a análise das métricas de avaliação nos diferentes cenários, apresentam resultados diferentes. A Regressão Logística é o classificador que melhor se comporta com a utilização da técnica de pré-processamento *stemming* e por outro lado o classificador Support *Vector Machines* obtém melhores resultados com a técnica de pré-processamento *lematização.*

Adicionalmente aos modelos aplicados anteriormente, efetuamos um conjunto de análises de forma a comparar os resultados obtidos quando utilizamos diferentes vetores e ao mesmo tempo diferentes técnicas de pré-processamento (*stemming* e lematização).

A avaliação dos diversos classificadores, com dois *vetorizadores* e com a técnica de pré-processamento *stemming* e com a *lematização*, resulta em 4 linhas por classificador, como podemos comprovar na tabela seguinte:

RL	
Lematização	
CountVectorizer	0,77
TFIDFVectorizer	0,77
Stemming	
CountVectorizer	0,83
TFIDFVectorizer	0,79

Figura 23 - RL c/ dois vetores, Stemming e lematização

SVM	
Lematização	
CountVectorizer	0,83
TFIDFVectorizer	0,8
Stemming	
CountVectorizer	0,78
TFIDFVectorizer	0,79

Figura 24- SVM c/ dois vetores Stemming e Lematização

Nas duas figuras anteriores apresentamos os melhores resultados obtidos. Destacando a Regressão Logística consegui um melhor no coeficiente *kappa*.

Por cada classificador, testamos dois vetores e por cada vetor testamos as duas técnicas de pré-processamento.

Na tabela seguinte, podemos encontrar todos os resultados obtidos para todos os classificadores.

Modelo	rocessament	Vetorizador	T. Exatidão	Kappa	ensibilidad	*F-measure*	T.Erro	Precisão
RL	Stemming	CountVectorizer	**0,83**	**1**	**0,81**	**0,81**	**0,18**	**0,85**
RL	Lematização	CountVectorizer	0,77	1	0,74	0,77	0,23	0,8
RL	Stemming	TFIDFVectorizer	**0,79**	**1**	**0,89**	**0,79**	**0,21**	**0,77**
RL	Lematização	TFIDFVectorizer	0,77	0,7	0,8	0,77	0,23	0,74
SVM	Stemming	CountVectorizer	0,78	0,59	0,8	0,78	0,21	0,77
SVM	Lematização	CountVectorizer	**0,83**	**0,66**	**0,8**	**0,83**	**0,17**	**0,86**
SVM	Stemming	TFIDFVectorizer	0,79	1	0,8	0,78	0,21	0,77
SVM	Lematização	TFIDFVectorizer	0,8	0,59	0,81	0,79	0,21	0,78
KNN	Stemming	CountVectorizer	0,64	0,3	0,84	0,69	0,36	0,6
KNN	Lematização	CountVectorizer	0,62	0,26	0,84	0,67	0,38	1
KNN	Stemming	TFIDFVectorizer	0,74	1	0,83	0,76	0,26	0,7
KNN	Lematização	TFIDFVectorizer	0,74	0,49	0,83	0,75	0,26	0,69
BernoulliNB	Stemming	CountVectorizer	0,8	0,61	0,86	0,81	0,19	0,77
BernoulliNB	Lematização	CountVectorizer	0,78	0,56	0,89	0,79	0,22	0,71
BernoulliNB	Stemming	TFIDFVectorizer	0,74	0,5	0,89	0,78	0,25	0,69
BernoulliNB	Lematização	TFIDFVectorizer	0,73	0,46	0,89	0,76	0,27	0,66
MultinomialNB	Stemming	CountVectorizer	0,78	0,5	0,86	0,8	0,22	0,74
MultinomialNB	Lematização	CountVectorizer	0,78	0,56	0,89	0,79	0,22	0,71
MultinomialNB	Stemming	TFIDFVectorizer	0,76	0,51	0,89	0,78	0,24	0,69
MultinomialNB	Lematização	TFIDFVectorizer	0,75	0,5	0,89	0,78	0,25	0,69
SGD	Stemming	CountVectorizer	0,81	0,53	0,75	0,78	0,22	0,81
SGD	Lematização	CountVectorizer	0,76	0,57	0,3	0,74	0,24	0,76
SGD	Stemming	TFIDFVectorizer	0,78	0,56	0,77	0,77	0,77	0,77
SGD	Lematização	TFIDFVectorizer	0,77	0,24	0,79	0,77	0,23	0,73
DT	Stemming	CountVectorizer	0,79	0,53	0,79	0,78	0,2	0,78
DT	Lematização	CountVectorizer	0,76	0,5	0,69	0,73	0,25	0,78
DT	Stemming	TFIDFVectorizer	0,72	0,44	0,73	0,72	0,28	0,71
DT	Lematização	TFIDFVectorizer	0,71	0,43	0,7	0,7	0,29	0,7

Tabela 34 - Comparação de todos os classificadores

Com a análise dos valores obtidos, verificamos que em média os classificadores *Support Vector Machine* e Regressão Logística são os que apresentam, sempre, a melhor taxa de exatidão

O *Support Vector Machine* apresenta melhores valores quando aplicamos a técnica de pré-processamento lematização juntamente com o *countvectorizer*, por outro lado a Regressão Logística apresenta melhores resultados quando empregamos a técnica de pré-processamento *stemming* simultaneamente com o *countvectorizer.*

No que diz respeito ao *TFIDFVectorizer, também* produz melhores resultados com os classificadores *Support Vector Machine e* Regressão Logística, apesar de não serem melhores quando aplicamos o *countvectorizer.*

Através da tabela 34 identificamos que, quanto à taxa de exatidão o modelo SVM, com lematização e com *countvectorizer* apresenta o maior valor, 83%, assim como, a

Regressão Logística, com o *stemming* e com *countvectorizer*, 83%. Por outro lado, o KNN, com lematização e com *countvectorizer* evidencia o menor valor (62%), isto resulta numa diferença de 21% entre estes.

No que diz respeito à precisão, verificamos que o modelo com maior valor é o SVM com lematização e com *countvectorizer* (86%), no entanto, o modelo com menor valor continua a ser o KNN, com stemming e com *countvectorizer (62%).*

No caso da sensibilidade *(recall)* obtivemos 3 classificadores com o maior valor (89%). O BernoulliNB lematização e com *countvectorizer* e com o vetor *TFIDFVectorizer,* tanto com a *lematização* como com o *stemming* obtiveram o mesmo resultado. *O MultinomialNB* também produziu o maior valor nesta métrica de avaliação, com recurso à *lematização* e com *countvectorizer*, também com o vetor *TFIDFVectorizer*, produziu o maior valor com as duas etapas de pré-processamento. Por fim, a Regressão logística com recurso ao *stemming* com *TFIDFVectorizer.*

Para a métrica de avaliação *f-measure* o que produziu melhor resultado foi o modelo SVM com lematização e com *countvectorizer (83%), por* outro lado o KNN voltou a ter o pior resultado (67%) com lematização e com *countvectorizer.*

A taxa de erro mais baixa pertence à SVM com lematização e com *countvectorizer,* com uma taxa de erro de 17%. O classificador SGD, obteve a taxa de erro mais alta (77%), com recurso ao Stemming e com o vetor *TFIDFVectorizer.*

Ao observar os resultados obtidos para o coeficiente kappa, podemos concluir que na maior parte dos classificadores obteve valores abaixo dos 60%. Apenas a Regressão logística e o KNN conseguirão 100% nesta métrica. Um dado curioso, a Regressão Logística em todos dos cenários, ou seja, tanto com o *countvectorizer* com *TFIDFVectorizer* e com a técnica de pré-processamento *(stemming)*, conseguiram 100% no coeficiente *kappa,* o que não se verificou com a aplicação do *TFIDFVectorizer e com* a utilização da técnica de lematização. Esta obteve 0.7 no coeficiente *kappa.*

Tendo em consideração os resultados das análises efetuadas, começamos por realçar a importância da aplicação da técnica de pré-processamento de remoção de *stopwords*. Foi evidente a diferença entre o número de palavras iniciais e finais.

No que diz respeito, à avaliação da qualidade dos sentimentos efetuada com os diversos classificadores, realçar o classificador de Regressão Logística.

Podemos concluir que o classificador de Regressão Logística apresentou o melhor resultado (83% na taxa de exatidão e com o *kappa* 1). Como referido anteriormente, quanto maior a proximidade do kappa=1, podemos confiar na taxa de exatidão conseguida.

Nesta dissertação foi elaborada uma investigação que abordava o impacto da inteligência artificial no negócio eletrónico. No início desta dissertação os obstáculos foram diversos, como por exemplo o não conhecimento sobre inteligência artificial, o que se revelou um grande desafio que penso ter sido superado com sucesso.

Na fase inicial desta dissertação fizemos um estudo sobre o estado da arte onde foi possível abordar diversos temas onde a inteligência artificial tem um grande impacto no negócio eletrónico e onde é notório que o futuro do negócio eletrónico passa pelo recurso a algoritmos de inteligência artificial.

A partir da informação recolhida, conseguimos aperceber-nos de que as redes sociais, *websites*, *fóruns* entre outras plataformas são ferramentas preciosas no negócio eletrónico. Os comentários nas redes socias podem demonstrar muito sobre um determinado produto e este pode influenciar outras pessoas na tomada de decisão sobre a compra. Foi a partir desta premissa, que consideramos relevante proceder a uma análise de sentimentos.

A Análise de Sentimentos, com recurso a comentários nas redes socias, possibilita-nos entender qual o sentimento presente no comentário, ou seja, se é positivo ou negativo relativamente a um produto ou serviço. Com recursos a técnicas de *Text Mining* conseguimos obter uma classificação dos documentos (comentários em redes sociais, *blogs*, entre outros).

A análise de sentimentos requer algum tempo de execução numa fase inicial onde se inclui a recolha dos dados e posterior preparação dos mesmos (pré-processamento dos dados recolhidos) e a aplicação de técnicas de *Machine Learning* para de forma automática se perceber o sentimento do utilizador. Os resultados obtidos mostraram que a aplicação de algoritmos de *text ming* aos comentários nas redes sociais é uma excelente ferramenta para avaliar se um produto está a ser bem recebido ou não pelo cliente. A opinião do cliente é uma importante ferramenta para conseguirmos retirar informação e entender as expetativas e opiniões do cliente. Deste modo, ao monitorizarmos estes processos permite-nos descobrir onde podemos melhorar o nosso negócio, bem como, um produto ou serviço.

Ao analisar a taxa de exatidão das diferentes combinações verificamos que o que apresentou melhores resultados foi a Regressão Logística e *o SVM* com a utilização de *countvectorizer*. No entanto, tendo em conta o valor do *kappa* (igual a 1) a RL foi o algoritmo que obteve melhores resultados.

Desta forma, podemos concluir que a análise de sentimentos é uma excelente ferramenta de análise relativamente aos sentimentos dos consumidores sobre um determinado produto ou serviço. permitindo-nos uma melhor tomada de decisão no futuro. A utilização destas tecnologias permite-nos obter uma vantagem competitiva e uma visão estratégica de modo a conseguirmos um objetivo crucial em qualquer negócio, a satisfação do cliente. É de salientar a importância da contínua análise das redes socias, pois, as mesmas estão em constante mutação bem como o comportamento do consumidor.

Com base nos dados apresentados, anteriormente, o classificador que evidenciou melhores resultados, com os nossos dados e as técnicas utilizadas, foi o de Regressão Logística. Após a conclusão da elaboração da dissertação obtivemos as evidências necessárias para concluir que a inteligência artificial tem impacto no negócio eletrónico, entendo que o processo de análise de sentimentos conduzido poderá facilitar as empresas efetuarem uma leitura e compreensão do comportamento de compra de atuais e potenciais consumidores.

Em suma, consideramos que uma análise similar à elaborada constitui uma excelente ferramenta para avaliar se um produto/ serviço é bem recebido ou não pelo cliente.

Limitações e Perspetivas futuras

A minha área de formação, Gestão, representou uma limitação na elaboração desta dissertação, providenciando-me poucas bases de trabalho prático útil nesta área.

A título de exemplo, o meu percurso de aprendizagem da língua de programação *Python*, com a qual nunca tinha trabalhado até então. Inicialmente, houve um processo de aprendizagem para me familiarizar com esta linguagem de programação. Contudo, penso que os objetivos definidos inicialmente foram alcançados. De certo modo, acredito que se tivesse conseguido aprofundar melhor este tema conseguiria obter melhores resultados. Em conclusão, se recomeçasse hoje este estudo, fundamentado nos conhecimentos já adquiridos, seria certamente mais elaborado.

Tratando-se de um tema com relevância em múltiplas áreas da tecnologia, e com uma aplicação cada vez mais presente em todas as áreas do nosso dia a dia, acredito que existe uma grande margem de investigação e criação à volta do mesmo. Esta é cada vez mais credibilizada pelos crescentes estudos, publicações e novas descobertas na área, oferecendo novas perspetivas e formas de implementação. Consideramos que esta

investigação é de suma importância para o negócio eletrónico, contribuindo assim para a sua continuidade e abrangência.

Como trabalho futuro, consideremos que seria interessante a realização de uma análise mais aprofundada, fundamentada com um maior volume de dados, como por exemplo um *dataset* que contenha um maior número de *reviews* para os resultados serem ainda mais fundamentados e conclusivos. Fazer uma comparação com trabalho já existentes e publicados, utilizar diferentes técnicas de pré-processamento de texto, uso de diferentes classificadores. Considero que a continuidade e abrangência desta investigação é de grande importância para o negócio eletrónico.

Ahlgren, M. (2022, abril 28). *Mais de 40 estatísticas do Twitter de 2022: Estatísticas, dados demográficos do usuário e fatos*. Website Rating. https://www.websiterating.com/pt/research/twitter-statistics/

Andonov, A., Dimitrov, G. P., & Totev, V. (2021). Impact of E-commerce on Business Performance. *TEM Journal*, *10*(4), 1558–1564. https://doi.org/10.18421/TEM104-09

Assery, N., Xiaohong, Y., Almalki, S., Kaushik, R., & Xiuli, Q. (2019). Comparing Learning-Based Methods for Identifying Disaster-Related Tweets. *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 1829–1836. https://doi.org/10.1109/ICMLA.2019.00295

Benevenuto, Fabrício, (Ribeiro), (Filipe), & (Matheus), (Araújo). (2015). *Métodos para Análise de Sentimentos em mídias sociais* (file:///C:/Users/claud/Downloads/19-Manuscrito%20de%20cap%C3%ADtulo-288-1-10-20190825.pdf).

Bertges, G., Moraes, L. A., & Zonovelli, B. (sem data). *CHATBOTS E SISTEMAS DE RECOMENDAÇÃO: DESAFIOS DA TECNOLOGIA PARA O ATENDIMENTO DE QUALIDADE NO E-COMMERCE*. 12.

Bullejos, M., Cabezas, D., Martín-Martín, M., & o Javier Alcalá, F. (2022). *A K-Nearest Neighbors Algorithm in Python for Visualizing the 3D Stratigraphic Architecture of the Llobregat River Delta in NE Spain*. 13.

Cappelli, C., ACM Digital Library, ACM Special Interest Group on Management Information Systems, & SIGAPP. (2016). *Proceedings of the XII Brazilian Symposium on Information Systems on Brazilian Symposium on Information Systems Information Systems in the Cloud Computing Era—Volume 1*. Brazilian Computer Society. http://dl.acm.org/citation.cfm?id=3021955

(Correia, Daniel). (2022). *Classificação de Dados Biológicos: Características e Classificadores*. https://comum.rcaap.pt/bitstream/10400.26/17234/1/Daniel-Joao-Correia.pdf

Cotter, S., Cordeiro, H., & Marques, G. (2022). Classification of Patients Diagnosed with ALS Through Spectral Parameters: Comparison between automatic classifiers and visual methos. *2022 17th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–6. https://doi.org/10.23919/CISTI54924.2022.9820415

Dash, R., McMurtrey, M., Rebman, C., & K. Kar, U. (2019). *Application of Artificial Intelligence in Automation of Supply Chain Management*.

Dastani, M., Chelak, A. M., Ziaei, S., & Delghandi, F. (2020). Identifying Emerging Trends in Scientific Texts Using TF-IDF Algorithm: A Case Study of Medical Librarianship and Information Articles. *Health Technology Assessment in Action*. https://doi.org/10.18502/htaa.v4i2.6231

Davenport, T. H., & Ronanki, R. (2018). *Artificial Intelligence for the Real World*. 10.

DeLone, W. H., & McLean, E. R. (2004). Measuring e-Commerce Success: Applying the DeLone & McLean Information Systems Success Model. *International Journal of Electronic Commerce*, *9*(1), 31–47. https://doi.org/10.1080/10864415.2004.11044317

Demircan, M., Seller, A., Abut, F., & Mehmet, F. A. (2021). Developing Turkish sentiment analysis models using machine learning and e-commerce data. *Elsevier B.V. on behalf of KeAi Communications Co. Ltd*. https://reader.elsevier.com/reader/sd/pii/S2666307421000231?token=91D7FFC715DDCD0A3F126CBADC0F4D09084AFBB9AC2FDCD709A1D513A63B541F756FA12E18A70EDB54586D0FEFD246A2&originRegion=eu-west-1&originCreation=20220822093810

Feinerer, I. (2008). *A text mining framework in R and its applications* [Doctoral, WU Vienna University of Economics and Business]. https://epub.wu.ac.at/1923/

Feldman, S. (1999). NLP Meets the Jabberwocky: Natural Language Processing in Information Retrieval', Online , 23:3. URL: http://www.onlineinc.com/onlinemag/OL1999/feldman5.html Feldman, S.(Jan 2000), 'The Answer Machine', Searcher , 8:1. URL: http://www.infotoday.com/search. *Proceedings of the Fifth Hong Kong Web Symposium*.

Fernandes de Avila, D., Dietrich Klug, W., Oreja, E. C., Martins Rodriguez, A., & do Amaral Martins Grimmler, J. (2022). Internet of Things E Inteligência Artificial Nos Meios Produtivos: INTERNET OF THINGS AND ARTIFICIAL INTELLIGENCE IN PRODUCTIVE MEANS. *Revista CIATEC-UPF*, *14*(2), 156–165. https://doi.org/10.5335/ciatec.v14i2.13789

Godbole, S., Bhattacharya, I., Gupta, A., & Verma, A. (2010). Building re-usable dictionary repositories for real-world text mining. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management - CIKM '10*, 1189. https://doi.org/10.1145/1871437.1871588

Gouveia, L. (2019). *Optimizing Facebook campaign's performance using Text Mining*.

Guerra, A. R. (2022, julho 28). *Amazon cresce e supera previsões em temporada de resultados desafiante*. Dinheiro Vivo. https://www.dinheirovivo.pt/empresas/tecnologia/amazon-cresce-e-supera-previsoes-em-temporada-de-resultados-desafiante-15055900.html

Kaczorowska-Spychalska, D. (2019). Chatbots in marketing. *Management*, *23*(1), 251–270.

Kaur, W., Balakrishnan, V., Rana, O., & Sinniah, A. (2019). Liking, sharing, commenting and reacting on Facebook: User behaviors' impact on sentiment intensity. *Telematics and Informatics*, *39*, 25–36. https://doi.org/10.1016/j.tele.2018.12.005

Khrais, L. T. (2020). Role of Artificial Intelligence in Shaping Consumer Demand in E-Commerce. *Future Internet*, *12*(12), Art. 12. https://doi.org/10.3390/fi12120226

Laudon, K. C., & Traver, C. G. (2007). *Copyright © 2007 Pearson Education, Inc.* 31.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*.

M. Mostafa, M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Elsevier Ltd.*

Matana, M. (2022). *A PROTEÇÃO DO CONSUMIDOR NO E COMMERCE*.

Matias, A. C. (2016). *A influencia da Inteligência Artifiicial no E-commerce—Uso dos chatbots*.

Mincov, B. M., Novakovski, T., Paula, K. J. S. de, Castro, G. C., Saganski, G. F., & Freire, M. H. de S. (2022). Processo de Validação de Tecnologia Educacional para o cuidado do paciente infanto juvenil oncológico submetido ao Transplante de Células-tronco Hematopoéticas: Revisão Integrativa. *Research, Society and Development*, *11*(11), Art. 11. https://doi.org/10.33448/rsd-v11i11.33832

Moreira, R. P. D. M. (2021). *Inteligência Artificial no Marketing Digital: Aceitação da utilização de Inteligência Artificial nas plataformas de comércio eletrónico*. https://comum.rcaap.pt/handle/10400.26/39527

Nakayama, M., & Wan, Y. (2018). *The cultural impact on social commerce: A sentiment analysis on Yelp ethnic restaurant reviews*. https://reader.elsevier.com/reader/sd/pii/S0378720617306225?token=D961458DD78A7C23E10A359FAFBA176596E17CD37B9F94890BCB7BC974ECAB684

AF91D88F6383F463BCEFD2C9AB9DAB6&originRegion=eu-west-1&originCreation=20220822104910

Nguyen, H., Veluchamy, A., Diop, M., & Iqbal, R. (2018). *Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches*. *1*(4), 23.

Prasetijo, A. B., Isnanto, R. R., Eridani, D., Soetrisno, Y. A. A., Arfan, M., & Sofwan, A. (2017). Hoax detection system on Indonesian news sites based on text classification using SVM and SGD. *2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, 45–49. https://doi.org/10.1109/ICITACEE.2017.8257673

Ribeiro, A. C., & Frazão, R. (sem data). *Quebra-cabeças Machine Learning: Como selecionar Use Cases, Algoritmos e Tecnologias?* 11.

Rodrigues, H. (2016). *Mestrado Integrado em Engenharia Informática e Computação*.

Rodrigues, M. (2019). *Text Mining for Social Networks Sentiment Analysis. Two case studies: "FIFA World Cup 2018" and "Cristiano Ronaldo signs for Juventus"*. Text Mining for Social Networks Sentiment Analysis. Two case studies: "FIFA World Cup 2018" and "Cristiano Ronaldo signs for Juventus"

Segall, R., Zhang, Q., & Cao, M. (2022). *Web-Based Text Mining of Hotel Customer Comments Using SAS® Text Miner and Megaputer Polyanalyst®*.

Shankar, V. (2018). How Artificial Intelligence (AI) is Reshaping Retailing. *Journal of Retailing*, *94*(4), vi–xi. https://doi.org/10.1016/S0022-4359(18)30076-9

Silva, P. (2020). *Classificação Automática de Documentos: Seleção customizada do classificador*.

Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019). Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification. *2019 International Conference*

on Automation, Computational and Technology Management (ICACTM), 593–596. https://doi.org/10.1109/ICACTM.2019.8776800

Tan, L., Li, M. Y., & Kok, S. (2020). E-Commerce Product Categorization via Machine Translation. *ACM Transactions on Management Information Systems*, *11*(3), 11:1-11:14. https://doi.org/10.1145/3382189

Turban, E. (2015). *Business intelligence and analytics: Systems for decision support* (Tenth edition). Pearson.

Turban, E., King, D. N., & Marques, A. S. (2004). *Comércio eletrônico*. Prentice Hall,.

Vel S., S. (2021). *Pre-Processing techniques of Text Mining using Computational Linguistics and Python Libraries SAKTHI VEL S.* 6.

Witten, I. H. (sem data). *Computer Science, University of Waikato, Hamilton, New Zealand email ihw@cs.waikato.ac.nz*. 23.

Yin, R. K. (2011). *Applications of Case Study Research*. SAGE.