DEPARTAMENTO DE ENGENHARIA E CIÊNCIAS DA COMPUTAÇÃO

MESTRADO EM ENGENHARIA INFORMÁTICA E DE TELECOMUNICAÇÕES

UNIVERSIDADE AUTÓNOMA DE LISBOA

"LUÍS DE CAMÕES"

CLASSIFICAÇÃO DE INFLUENCERS DE MERCADOS FINANCEIROS EM REDES SOCIAIS

Dissertação para a obtenção do grau de Mestre em Engenharia Informática e de Telecomunicações

Autora: Inês Gonçalves Saúde de Almeida

Orientador: Professor Doutor Mário Pedro Guerreiro Marques da Silva

Coorientador: Professor Doutor André Miguel Guedelha Sabino

Número da candidata: 30001099

Maio de 2022

Lisboa

(Página em Branco)

Dedicatória

Dedico esta dissertação ao meu pai, por sempre acreditar em mim e por todo o apoio durante esta caminhada tão importante para mim, quer a nível pessoal como profissional. Eu nunca serei capaz de agradecer o suficiente por isso.

Agradecimentos

Quero, antes de mais, expressar o meu agradecimento a todos aqueles que dum modo ou de outro, contribuíram para a realização deste trabalho. Ainda que correndo o risco de injustamente omitir algumas pessoas/entidades cujo contributo foi relevante, não posso deixar de salientar o apoio recebido sem reservas:

- Do meu orientador Professor Doutor Mário Marques da Silva e do meu coorientador Professor Doutor André Sabino por todo o apoio, dedicação, horas despendidas, ensinamentos administrados e confiança depositada;

- Dos meus pais, pela dedicação, incentivo e pela possibilidade que me deram de poder ter uma formação superior.

- Da minha avó Natália por todo o incentivo que me deu para estudar e seguir os meus sonhos;

- De todos os meus amigos, pelo apoio durante a realização deste trabalho.

Epígrafe

“Uma criança, um professor, um livro e uma caneta podem mudar o mundo!”

Malala Yousafzai

Resumo

Atualmente, as redes sociais fazem parte do nosso quotidiano. É nestas que os utilizadores expressam as suas opiniões através de publicações, comentários e gostos/reações. As informações publicadas pelos utilizadores podem ter impacto na sua comunidade virtual, e assim nasce o conceito de *influencer* (i.e., influenciador). Normalmente cada *influencer* foca-se num determinado um assunto (como moda, saúde, bem-estar, beleza, finanças etc.) nas redes sociais sendo que a sua opinião impacta os utilizadores que a leem. Sendo um conceito que abrange diversas áreas, também existem *influencers* no mercado financeiro sendo estes perfis de utilizadores cuja análise de sentimento do conteúdo produzido está correlacionada com a volatilidade das ações de uma empresa. Os utilizadores que investem na Bolsa de valores utilizam a rede social Twitter como principal fonte para exprimir as suas opiniões sobre as ações. Os *tweets* realizados expressam um sentimento "Positivo", "Negativo" ou "Neutro" sobre aquela ação. Com esta dissertação pretende-se identificar o perfil de um *influencer* assim como analisar o impacto dos seus *tweets* no valor da ação em estudo. Para a descoberta dos valores que melhor descrevem o perfil de *influencer* para o mercado financeiro recorreu-se ao Algoritmo genético.

Sendo este tema foco de várias pesquisas nesta dissertação também foi produzida uma comparação entre técnicas de análise de sentimento, aplicadas ao conteúdo de redes sociais, produzido por potenciais *influencers*. Os métodos de análise de sentimento utilizados foram SVM, *Neural Network*, *Naive Bayes*, kNN (*k-nearest neighbors).* Após a análise dos resultados o método com o melhor resultado foi a *Neural Network,* com exatidão (*accuracy*) de 99,0 % para os *tweets* em estudo. O resultado da análise de sentimento produzida pela *Neural Network* foi utilizado como parâmetro para análise dos *tweets* dos *influencers*. Mas antes desta análise foi necessário detetar os parâmetros e respetivos valores que definem um utilizador *influencer*. Para essa deteção recorreu-se ao Algoritmo genético. Os resultados obtidos foram interessantes, já que ao realizar um caso estudo com utilizadores que constituem a solução obtida pelo Algoritmo genético é possível verificar o sentimento expresso nos *tweets* dos utilizadores estão de acordo com a alteração do valor da cotação da ação para estratégias de curta duração.

Palavras-chave: Abordagem Híbrida; Análise de sentimento; *Influencers*; *Lexicon-based*; *Machine Learning*; Mercado financeiro; Twitter;

Abstract

Nowadays, social networks are part of our daily lives, enabling users to express their opinions through publications, comments, and likes/reactions. The information published by users can have an impact on their virtual community, and thus the concept of *influencer* becomes relevant. Generally, a social network influencer focuses on a particular subject (such as fashion, health, well-being, beauty, finances, etc.), and their opinion impacts the community of users who share the same interests. The concept of social network influencer may also be explored in the financial markets, considering user profiles whose expressed opinions align with the performance of a company's shares. Several users who invest in the stock exchange use the social network Twitter as the main platform to express their opinions on stocks. The tweets express either a positive, negative, or neutral feeling about a stock. This thesis proposes to study and describe the profile of an influencer, as well as analyze the correlation between their tweets and the performance of a stock. We use a genetic algorithm to best describe an influencer's profile, selecting the best fit from a set of features.

We provide a comparison between sentiment analysis techniques, applied of social network content, produced by potential influencers. We study several sentiment analyses approaches, namely using Support Vector Machines, Neural Networks, Naive Bayes, and K-nearest neighbors. Results show that the method with the best performance is the Neural Network, with an accuracy of 99.0% for the tweets under study. This outcome is used as a parameter for analyzing the influencers' tweets.

The approach was successfully validated, showing that the integration of a genetic algorithm for influencer detection with a sentiment analysis technique based on Neuronal Networks, it is possible to identify users with social network content sentiment score correlated with variations in a share price.

Keywords: Financial market; Hybrid Approach; Influencers; Lexicon-Based; Machine Learning; Sentiment analysis; Twitter.

Índice

Dedicatória **3**

Agradecimentos **4**

Epígrafe **5**

Resumo **6**

Abstract **7**

Índice **8**

Lista de Quadros **11**

Lista de Figuras **12**

Lista de Abreviaturas **13**

Lista de Siglas e Acrónimos **14**

1 Introdução **15**

1.1 Objetivos 15

1.1.1 Objetivos específicos **15**

1.2 Caracterização do problema 16

1.3 Justificação 16

1.3.1 Justificação Teórica **17**

1.4 Fundamentação Teórica 17

1.5 Estrutura do documento 18

2 Revisão de literatura **19**

2.1 Twitter 20

2.2 Análise de sentimento 20

2.3 Aplicação de Análise de Sentimento a tweets 25

2.4 Utilização do Twitter para previsões no Mercado Financeiro 26

2.5 *Influencers* no Twitter 28

2.6 *Influencers* do mercado financeiro no Twitter 29

2.7 Algoritmo Genético 30

2.8 Síntese do capítulo 32
3 Metodologia 34
3.1 Arquitetura do Sistema 34
3.2 Módulo de recolha e processamento de dados 35
3.2.1 Obtenção dos *tweets* 35
3.2.2 Pré-processamento dos *tweets* 36
3.2.3 Obtenção de informação dos utilizadores 37
3.2.4 Obtenção do valor *Adjusted Closing Price* (ACP) da ação 38
3.3 Módulo de análise de sentimento 39
***3.3.1 Machine Learning* 40**
3.3.2 Rotulação do conjunto de treino – abordagem híbrida 40
3.3.3 Justificação para a utilização do módulo *Python* VADER 40
3.3.3.1 Aplicação dos métodos 41
3.4 Módulo de deteção de *influencers* 42
3.4.1 Representação dos indivíduos 42
3.4.2 Definição da função de avaliação: função de *fitness* 43
3.4.3 Processo de seleção 44
***3.4.4* Processo de *Crossover* 44**
3.4.5 Mutação 45
3.4.6 Resumo dos parâmetros do Algoritmo genético 45
3.5 Síntese do capítulo 46
4 Validação do sistema 47
4.1 Recolha e processamento dos dados 47
4.1.1 Obtenção e pré-processamento de *tweets* 47
4.1.2 Obtenção de informação dos utilizadores 48
4.2 Obtenção de informações de mercado 49
4.3 Módulo de análise de sentimento 49

4.3.1 Criação dos conjuntos de teste e treino 49

4.3.2 Rotulação do conjunto de treino – abordagem híbrida 50

4.3.3 Aplicação dos métodos de *Machine Learning* 50

4.3.4 Análise do resultado da análise de sentimento..... 53

4.4 Módulo de deteção de *influencers* 54

4.5 Síntese do capítulo..... 55

5 Caso de estudo 56

5.1 Validação dos resultados para o *Rank* 1 56

5.1.1 Estratégia de curta duração 57

5.1.2 Estratégia de longa duração 58

5.2 Validação dos resultados para a característica Número de lista públicas 59

5.2.1 Estratégia de curta duração 59

5.2.1 Estratégia de longa duração 60

5.3 Síntese do capítulo..... 61

6 Conclusões 62

7 Sugestões para trabalhos futuros 63

Bibliografia..... 64

Anexo 01 – Repositório GitHub 67

Lista de Quadros

Tabela 1 - Resumo do capítulo 2 Fonte: elaboração própria 33
Tabela 2 - Comparação entre sistema da dissertação base e esta dissertação Fonte: elaboração própria 34
Tabela 3 - Recursos utilizados no módulo de recolha e processamento de dados: dissertação base vs. Esta dissertação Fonte: elaboração própria 39
Tabela 4 - Abordagens de análise de sentimento aplicadas: dissertação base vs. Esta dissertação Fonte: elaboração própria 40
Tabela 5 - Número total de tweets recolhidos Fonte: elaboração própria 48
Tabela 6 - Número total de utilizadores selecionados para a simulação Fonte: elaboração própria 48
Tabela 7 - Número de tweets rotulados como “Positivo”, “Negativo” ou “Neutro” Fonte: elaboração própria 50
Tabela 8 - Fases de aplicação dos métodos de Machine Learning e respetivos ícones Fonte: elaboração própria 52
Tabela 9 - Resultados obtidos pelo Algoritmo genético Fonte: elaboração própria 54
Tabela 10 - Análise do impacto dos tweets do utilizador szaman Fonte: elaboração própria . 58
Tabela 11 - Análise do impacto dos tweets do utilizador jonahlupton Fonte: elaboração própria 58
Tabela 12 - Análise do impacto dos tweets do utilizador BP_Swing_Trader Fonte: elaboração própria 60

Lista de Figuras

Figura 1 - Relação entre o tema da dissertação e os principais tópicos deste capítulo Fonte: elaboração própria 19
Figura 2 - Comparação entre Aprendizagem por reforço, Aprendizagem supervisionada e aprendizagem não supervisionada [9] 22
Figura 3 - Exemplo de SVM linear e não linear [13] 23
Figura 4 - Métodos principais da análise de sentimento [12] 23
Figura 5 - Etapas do Algoritmo genético Fonte: elaboração própria 31
Figura 6 - Fases do módulo de recolha e processamento dos dados Fonte: elaboração própria 35
Figura 7 - Fases do pré-processamento dos tweets Fonte: elaboração própria 37
Figura 8 - Diferença entre o cromossoma desta dissertação vs. dissertação base [2] 38
Figura 9 - Representação dos Ranks utilizados nesta dissertação Fonte: elaboração própria .. 43
Figura 10 - Representação do cromossoma aplicado nesta dissertação Fonte: elaboração própria 43
Figura 11 – Exemplo do processo de seleção desta dissertação Fonte: elaboração própria 44
Figura 12 - Processo de Crossover desta dissertação Fonte: elaboração própria 45
Figura 13 - Detalhes dos conjuntos de treino e de teste Fonte: elaboração própria 50
Figura 14 - Aplicação dos métodos de Machine Learning através do Orange Fonte: elaboração própria 52
Figura 15 - Resultados dos métodos de Machine Learning Fonte: elaboração própria 53
Figura 16 - Matriz de confusão de Neural Network Fonte: elaboração própria 54

Lista de Abreviaturas

API	*Application Programming Interface*
ACP	*Adjusted Closing PriceComma-separated values*
CSV	*Comma-separated values*
GOT	*Get Old Tweets*
MF-DCCA	*Multifractal Detrended Cross-Correlation Analysis*
ROC	*Receiver operating characteristic*
SVM	*Support-vector Machine*
VADER	*Valence Aware Dictionary for sEntiment Reasoning*

Lista de Siglas e Acrónimos

ACM	*Association for Computing Machinery*
DIJA	*Dow Jones Industrial Average*
EUA	Estados Unidos da América
NASDAQ	*National Association of Securities Dealers Automated Quotations*
NYSE	*New York Stock Exchange*

1 Introdução

Este tema é alvo de estudo de várias pesquisas, cujo resultado é obtido através da análise e aplicação de um único método de análise de sentimento. Uma vez que existem várias alternativas para realizar a análise de sentimento [1], pretende-se testar os métodos principais da análise de sentimento (*Lexicon-based e Machine Learning*) bem como a sua combinação de forma a analisar a eficácia de cada um destes métodos.

1.1 Objetivos

O objetivo geral desta dissertação é classificar *influencers* de mercados financeiros em redes sociais.

É importante mencionar que este tema já foi alvo de estudos de diversas pesquisas, tendo sido obtidas soluções que ainda podem ser melhoradas [2], [3], [4]. Nestas publicações apenas é aplicado um método de análise de sentimento, contudo através de investigação foi possível detetar que a combinação de vários métodos de análise de sentimento poderá obter um melhor resultado.

Neste projeto será realizada a classificação de *influencers* através dos dois principais métodos (*Lexicon-based* e *Machine Learning*) e da combinação de ambos, de forma a verificar qual o método que fornece melhor resultado.

1.1.1 Objetivos específicos

- Identificar o conjunto de utilizadores alvo de classificação (amostra);
- Analisar o sentimento dos *tweets* da amostra que mencionam ações do Índice *Standard & Poor's* 500 (Índice composto pelos quinhentos ativos cotados nas bolsas de *New York Stock Exchange* (NYSE) ou *National Association of Securities Dealers Automated Quotations* (NASDAQ), com base em vários métodos de análise de sentimento;
- Analisar o resultado produzido pelos diversos métodos de análise de sentimento;
- Classificar cada utilizador da amostra como *influencer* e não *influencer*.

Após o cumprimento dos objetivos específicos será realizada uma análise do impacto dos *influencers* encontrados na volatilidade (i.e., desvio padrão da série temporal de preços) das ações e assim melhorar a experiência dos investidores com a aposta em ferramentas que os ajudem no dia a dia.

1.2 Caracterização do problema

Problema: Qual o modelo de análise de sentimento mais eficaz para classificação de *influencers* de mercados financeiros em redes sociais?

A deteção de *influencers* relaciona o comportamento de ações no mercado financeiro com o conteúdo publicado por utilizadores em redes sociais. Pretende-se analisar o valor explicativo que este conteúdo tem sobre o comportamento do mercado, particularmente qual a melhor estratégia para interpretar o sentimento associado ao conteúdo com o objetivo de o utilizar num modelo de classificação

O protocolo de investigação a aplicar envolve a coleção sistemática de informação de redes sociais e mercados financeiros, e a aplicação de modelos de avaliação de sentimento sobre o conteúdo social, que servirá para alimentar modelos de classificação treinados para detetar *influencers*.

Uma contribuição deste trabalho é a avaliação de vários modelos de análise de sentimento. A eficácia destes modelos de análise de sentimento envolve a sua aplicação num modelo de classificação, cuja validação será realizada sobre dados históricos, tentando detetar *influencers* conhecidos com base no conteúdo por estes publicado no passado.

A deteção de *influencers* permite obter um sinal sobre o qual se podem construir modelos de investimento.

Existem várias alternativas para realizar a análise de sentimento. Pretende-se testar vários métodos baseados em aprendizagem automática, comparando-os com métodos baseados em análise lexicográfica, e com métodos que combinam várias abordagens [1].

A hipótese que se pretende avaliar propõe que métodos que combinam princípios de aprendizagem automática com análise lexicográfica têm uma prestação superior às abordagens de apenas uma das áreas.

1.3 Justificação

O mercado tem, conhecidamente, compreensão difícil. Para um investidor é necessário identificar claramente o que está à procura, qual o resultado que pretende obter e, talvez mais importante, onde está a errar, relembrando sempre da máxima de Peter Drucker: “Não pode ser gerido o que não pode ser medido”.

Do ponto de vista prático, o objetivo principal desta investigação é contribuir para o apoio da tomada de decisão dos investidores através do fornecimento de uma base de

informação fundamentada que permita apoiar investidores iniciantes a tomar decisões mais informadas, e enriqueça as fontes de informação dos investidores mais experientes.

Desta forma, os agentes do mercado, quer investidores, quer especuladores, encontram nesta dissertação bases científicas que contribuem para atingir os seus objetivos, ou seja, obter rentabilidade positiva em investimentos no mercado bolsista, ou reduzir eventuais perdas.

1.3.1 Justificação Teórica

Sendo um tema bastante abordado em várias publicações e alvo de diversas pesquisas [2], [3], [4], o proposto nesta dissertação é cooperar para um melhoramento das investigações realizadas através das seguintes principais contribuições:

- Análise sentimental dos *tweets* dos utilizadores com base na combinação dos principais métodos de análise de sentimento (*Lexicon-based* e *Machine Learning*);
- Avaliação crítica dos principais métodos de análise de sentimento aplicados aos *tweets*;
- Criação de um sistema automático de identificação de *influencers* para uma determinada ação.

1.4 Fundamentação Teórica

A classificação de *influencers* de mercados financeiros em redes socias têm sido alvo de implementação e pesquisa de diversas publicações incluindo dissertações. Em relação a esta última destaca-se a dissertação [2], onde foi implementado um sistema de classificação ternária ("Positivo", "Negativo" ou "Neutro") do sentimento de *tweets* e a criação de um sistema de identificação de *influencers* com base no algoritmo *Naive Bayes* e Algoritmos genéticos.

Como principal pesquisa publicada sobre o tema ainda se destaca [3] onde se conclui através do sistema desenvolvido que a maioria dos influenciadores encontrados são contas oficiais de empresas ou de media de notícias.

O objetivo desta dissertação é contribuir para um melhoramento das investigações já efetuadas, através da implementação de um sistema de classificação de *influencers* no mercado financeiro que tem por base a aplicação, análise e comparação dos principais métodos de análise de sentimento.

A análise de sentimento divide-se em dois métodos principais: *Machine Learning* e *Lexicon-based*. Os métodos de análise lexicográfica (*Lexicon-based*) têm como vantagem a sua simplicidade, flexibilidade para diferentes linguagens e rapidez de análise.

1.5 Estrutura do documento

Este documento é composto 7 capítulos:

- Capítulo 1 – Introdução: Neste capítulo é apresentado os objetivos desta dissertação assim como a caraterização do problema. Ainda é apresentado neste capítulo uma justificação prática e teórica para escolha deste tema.
- Capítulo 2 – Revisão da literatura: É analisado e discutido o trabalho anterior relacionado com os conceitos abordados nesta dissertação, dos quais se destacam:
    - A rede social Twitter;
    - Os métodos de análise de sentimento;
    - A aplicação de análise de Sentimento a *tweets*;
    - A utilização do Twitter para previsões no Mercado Financeiro;
    - *Influencers* no Twitter;
    - *Influencers* do mercado financeiro no Twitter.
- Capítulo 3 – Metodologia: É definido a arquitetura do sistema e os seus principais módulos. Também são descritas as fases principais para implementação do sistema proposto.
- Capítulo 4 – Validação do Sistema: Simulação do sistema proposto através da sua implementação com dados reais. Também são avaliados os resultados produzidos.
- Capítulo 5 – Caso de estudo: Análise dos resultados obtidos no Capítulo 4.
- Capítulo 6 – Conclusões: Análise de todo o trabalho realização e proposta de trabalho futuro relevante para a investigação científica nesta área.
- Capítulo 7 – Sugestões para trabalho futuro: Indicação de relevantes melhorias futuras para a continuação do trabalho realizado nesta dissertação.

2 Revisão de literatura

A classificação de *influencers* de mercados financeiros em redes sociais, envolve um conjunto de conceitos fundamentais para a sua implementação. Este capítulo, fornece uma visão geral do trabalho relevante já desenvolvido por vários autores sobre cada conceito fundamental para o tema desta dissertação, nomeadamente o Twitter. Esta rede social é utilizada nesta dissertação como fonte principal de dados. Neste capítulo é descrito o seu funcionamento.

Outro conceito fundamental é a análise de sentimento. Esta abordagem será utilizada para análise dos *tweets* produzidos por cada *influencer* encontrado. O seu funcionamento e principais métodos encontram-se apresentados e descritos neste capítulo. Sendo o Twitter e a análise de sentimento dois dos conceitos essenciais, neste capítulo é mencionado trabalho relevante de vários autores sobre a aplicação da análise de sentimento ao conteúdo produzido na rede social Twitter.

Para terminar, são apresentadas neste capítulo investigações realizadas por diversos autores sobre a utilização do Twitter para a classificação de *influencers* nos vários setores de atividade, com ênfase no mercado financeiro.

Face ao exposto, na Figura 1 encontra-se esquematizada a relação entre os principais conceitos desta dissertação (destacados a azul) e os tópicos mencionados neste capítulo (destacados a negrito).

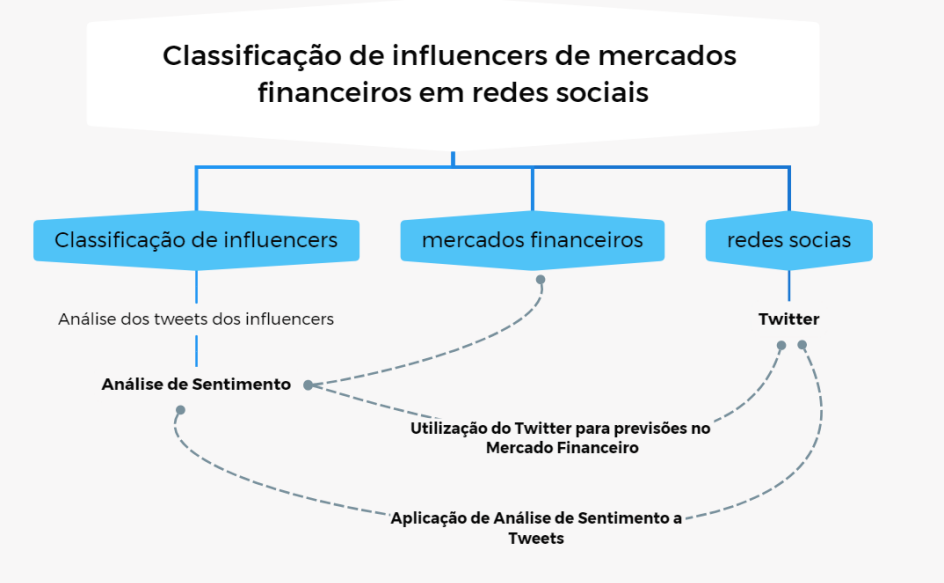


Figura 1 - Relação entre o tema da dissertação e os principais tópicos deste capítulo
Fonte: elaboração própria

2.1 Twitter

O Twitter é uma rede social que possibilita aos seus utilizadores a partilha, de forma gratuita, de informações e opiniões sobre política, personalidades, marcas, empresas, produtos, eventos, entre outros [1].

As mensagens publicadas pelos utilizadores são designadas de *tweets*, estas mensagens podem incluir "@NomeUtilizador" o que significa que o tweet é uma resposta a um determinado utilizador. O utilizador também pode encaminhar um tweet de outro utilizador, a mensagem resultado é designada de *retweet*. No texto do *tweet* pode ser introduzido *tags* como "#AAPL", designadas de hashtags [1].

O texto de cada *tweet* tem um limite de caracteres. Até 2017 os *tweets* tinham um limite de 140 caracteres [1]. Atualmente, alterou o comprimento máximo permitido de um *tweet* é de 280 caracteres [5]. Um estudo concluiu que o comprimento médio de um *tweet* é de 14 caracteres, o que é um resultado muito inferior quando comparado ao comprimento médio de uma frase (78 caracteres) [6]. Esta limitação no comprimento das mensagens (*tweet*) partilhadas pelos utilizadores é um dos fundamentais motivos para que o Twitter seja uma das redes principais para análise de conteúdo.

Z. Drus e H. Khalid [1] realizaram uma análise de artigos relacionados com análise de sentimento publicados entre 2014 e 2019 em base de dados de pesquisa como a *Association for Computing Machinery* (ACM), *Emerald Insight*, *Institute of Electrical and Electronics Engineers Xplore* (IEEE Xplore), *Science Direct* e *Scopus*. Concluiu-se que cerca de 85% dos artigos examinados utilizam como fonte de dados o Twitter [1]. Para complementar, destaca-se ainda como principal motivo para a seleção do Twitter como fonte de pesquisa de análise de conteúdo a facilidade de recolha de milhões de mensagens (*tweets*), através da sua *Application Programming Interface* (API). Neste estudo ainda é destacada a disponibilidade, acessibilidade e riqueza do conteúdo da rede social Twitter [1]. Esta dissertação utilizará o Twitter como fonte principal de dados.

2.2 Análise de sentimento

Um dos métodos de análise do conteúdo de um texto ou mensagem é designado de análise de sentimento. Define-se sentimento, como uma emoção pessoal que pode ser positiva ou negativa [6]. Análise de sentimento é um método de extrair subjetividade e polaridade de um texto, podendo ser classificado como: "Positivo", "Negativo" ou "Neutro" [7]. As duas abordagens principais da análise de sentimento são *Lexicon-based* e *Machine Learning*.

A primeira abordagem de classificação é não supervisionada e envolve o cálculo da orientação de um documento a partir de orientação semântica de palavras ou frases presentes no documento [7]. Geralmente, nesta abordagem recorre-se a dicionários de palavras anotadas com a orientação semântica da palavra ou polaridade. Estes dicionários podem ser criados de forma automática ou manual [7].

A abordagem *Lexicon-based* foi utilizada na pesquisa realizada por M. Taboada, J. Brooke, M. Tofiloski, K. Voll, e M. Stede [7], demonstrando que esta abordagem apresenta bom desempenho. Contudo, possui limitações em relação aos dicionários, como os dicionários são criados ou automaticamente ou manualmente, a sua confiabilidade é limitada [7].

Tom Mitchel define *Machine Learning*, como um programa de computador que aprende com a experiência sobre um conjunto tarefas, cujo desempenho aumenta com a experiência [9], [10].

Geralmente, *Machine Learning* é definido como uma aprendizagem supervisionada, sendo necessário um conjunto de treino. Como exemplo, considere-se que se pretende classificar um conjunto de espécies de plantas como Setosa, Versicolor e Virgínica com base no comprimento da sépala, largura da sépala, comprimento da pétala e largura da pétala [9].

Numa aprendizagem supervisionada é necessário que previamente seja formado um conjunto de treino, que contem plantas já previamente rotuladas com as respetivas classes (Setosa, Vesicolor ou Virgínica) [9]. Com base na classificação realizada para o conjunto de treino, é possível classificar o conjunto de teste (dados que não foram previamente classificados).

Em resumo, na aprendizagem supervisionada, cada exemplo de entrada (características dos dados) no conjunto de treino tem um conjunto de destinos de saída (categoria dos dados), e o objetivo é aprender o mapeamento da entrada e saída [11].

Para além da aprendizagem supervisionada, também existe a possibilidade de uma aprendizagem não supervisionada em *Machine Learning*. Numa aprendizagem não supervisionada, não é necessário um conjunto de treino, com dados previamente rotulados por classes e também não existe regras predefinidas para categorizar as classes [11].

No exemplo de classificação das plantas, se for utilizada aprendizagem não supervisionada as classes das espécies de plantas são desconhecidas, apenas é conhecido o número de classes (neste caso são 3). Com base nas características em comum os dados serão agrupados em categorias, cada categoria corresponderá a uma classe de plantas [11].

Para além da aprendizagem supervisionada e não supervisionada, também existe em *Machine Learning* a aprendizagem por reforço (Ver Figura 2). Neste tipo de aprendizagem o

sistema ou agente aprende como interagir com o ambiente. O sistema desconhece qual a melhor ação a realizar [11]. A título de exemplo pode-se considerar que em vez do sistema aprender com um professor que lhe vai indicando o que fazer em cada etapa, é como se o sistema aprendesse com um crítico que ocasionalmente vai indicando se ação está correta ou errada através de recompensas ou punições [11] .

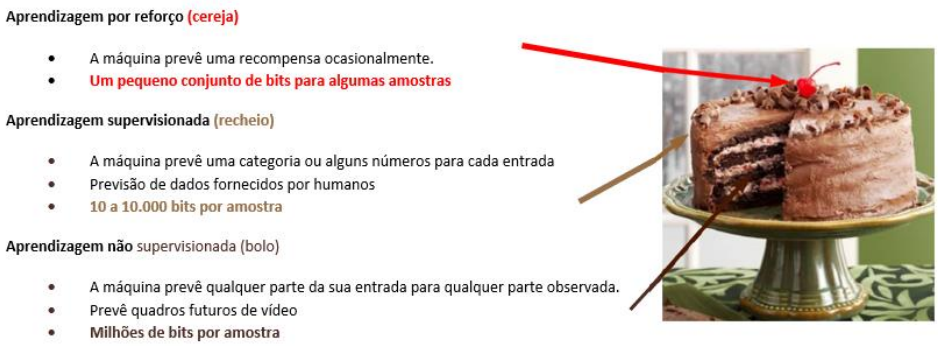


Figura 2 - Comparação entre Aprendizagem por reforço, Aprendizagem supervisionada e aprendizagem não supervisionada [9]

Nesta dissertação recorre-se a *Machine Learning* como uma abordagem de classificação supervisionada que envolve a construção de classificadores, a partir de dados previamente rotulados (i.e., conjunto de treino) [7].

Após pesquisa, foi possível encontrar diversas pesquisas que recorrem a métodos de *Machine Learning* para realizar análise de sentimento, como é o exemplo do trabalho de B. Pang, L. Lee e S. Vaithyanathan [11]. Nesta análise de sentimento recorreu-se às opiniões sobre um conjunto de filmes como fonte de dados para aplicação de métodos de *Machine Learning* [11]. Com esta pesquisa concluiu-se que os resultados produzidos pelas técnicas de *Machine Learning* superam as *human-produced baselines.* Os algoritmos de *Machine Learning* utilizados nesta pesquisa foram: *Naive Bayes*, *Maximum Entropy Classification* (ME) e *Support Vetor Machine* (SVM) [11].

O algoritmo *Naive Bayes* calcula a probabilidade posterior de uma classe, com base na distribuição das palavras no documento [12]. A metodologia ME converte conjuntos de dados rotulados em vetores através de descodificação. Este vetor é utilizado para calcular pesos para cada característica que pode então ser combinada para determinar o mais provável rótulo para um conjunto de recursos [12]. A SVM é um conjunto de métodos de aprendizagem supervisionada utilizados para classificação, regressão e deteção de *outliers* [13].

A SVM é eficaz em espaços de alta dimensão e como utiliza um subconjunto de pontos de treino na função de decisão (vetores de suporte), também é eficiente em termos de memória. Em termos de classificação a SVM pode separar as diferentes classes através de separadores lineares (SVM linear) ou separadores não lineares (SVM não linear) [13] (Ver Figura 3).

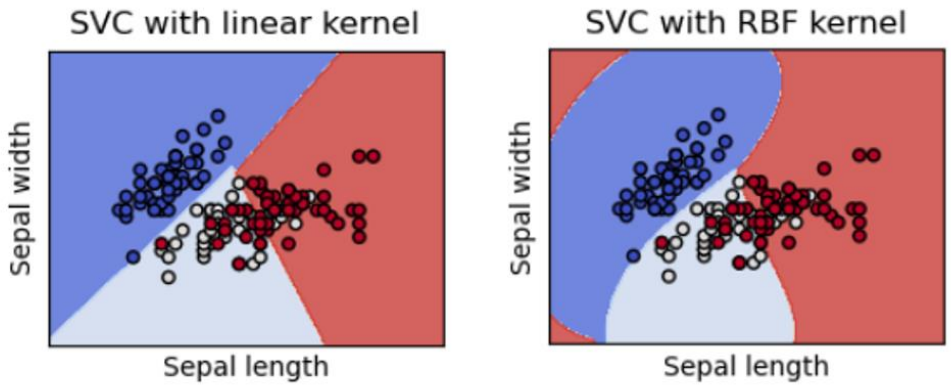


Figura 3 - Exemplo de SVM linear e não linear [13]

Nesta dissertação o objetivo da SVM é determinar separadores lineares no espaço de pesquisa que melhor pode separar as diferentes classes [12], sendo utilizada uma SVM linear.

As principais abordagens de análise de sentimento, englobam mais métodos para além dos descritos anteriormente (Ver Figura 4). Deste conjunto, destacam-se dentro de abordagem *Machine Learning*, os métodos *Naive Bayes*, *Neural Network* e SVM. A abordagem *Lexicon-based* é composta pelos métodos *Dictionary-based* e pelo método *Corpus-based Aproach* [12].

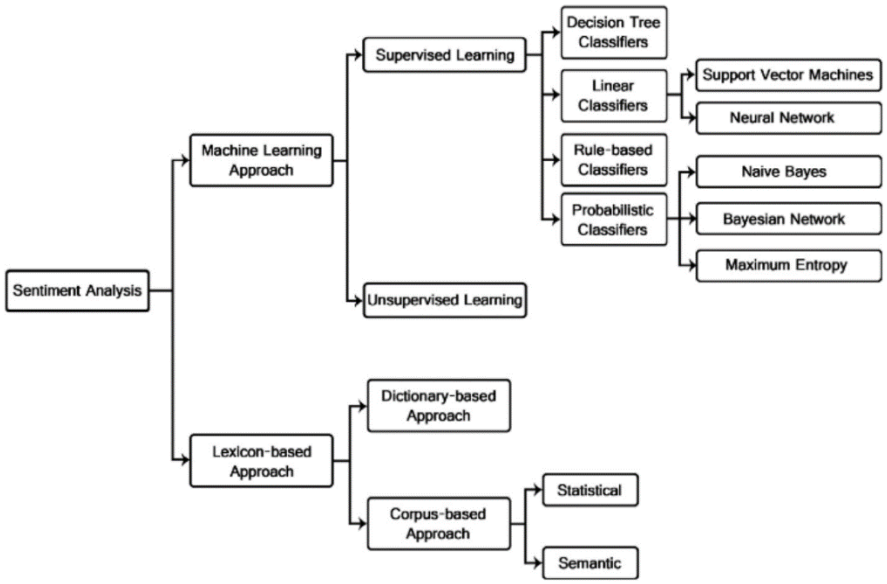


Figura 4 - Métodos principais da análise de sentimento [12]

Nesta dissertação será aplicada a abordagem *Machine Learning* através da utilização e análise dos seguintes métodos:

- SVM;
- *Neural Network* (M*ulti-layer Perceptron* (MPL))*;*
- *Naive Bayes;*
- *k-nearest neighbors* (kNN).

A SVM e o algoritmo *Naive Bayes* já foram definidos anteriormente neste capítulo, sendo também, importante definir *Neural Network* e kNN.

Nesta dissertação recorre-se à plataforma *Orange*[1] (versão 3.26.0) para a utilização dos métodos referidos anteriormente e no caso da *Neural Network*, esta plataforma utiliza o algoritmo *Multi-layer Perceptron* (MPL) sendo este o algoritmo que será aqui apresentado. MPL é um algoritmo que aprende com base num conjunto de treino [14]. Através de um conjunto de dados e de um rótulo este algoritmo consegue aprender aproximações de funções não lineares [14].

Para além do MPL, esta dissertação também utiliza o método *k-nearest neighbors*. Este pode ser supervisionado ou não supervisionado, sendo que nesta dissertação será utilizado a abordagem supervisionada [15]. Para este algoritmo é definido um ponto, que corresponde ao novo elemento que se pretende classificar e através dos elementos já classificados no conjunto de treino é medido a distância entre cada um deles e este novo ponto. Em resumo, os novos elementos são classificados de acordo com a classificação dos seus vizinhos [15].

Em relação, à abordagem de *Lexicon-based* será utilizada o método *Dictionary-based* implementado pelo módulo *Python* designado de VADER[2]. Neste tipo de abordagem recorre-se a um conjunto de palavras previamente rotuladas de acordo com sua orientação semântica como “Positivo”, “Negativo” ou “Neutro” para determinar o sentimento da palavra/frase em análise [12].

Existem também sistemas que utilizam as duas abordagens (*Machine Learning* e *Lexicon-based*) em conjunto (abordagem híbrida). Como por exemplo o sistema *OpinionFinder* [16] que identifica automaticamente quando opiniões, sentimentos, especulações estão presentes no texto [16]. Recorreu-se à abordagem *Lexicon-based* para a identificação de dados de treino a serem utilizado na abordagem de *Machine Learning* escolhida para a classificação

[1] Orange https://orangedatamining.com/
[2] VADER https://pypi.org/project/vaderSentiment/

de subjetividade do texto [16]. Através da avaliação dos resultados concluiu-se que a combinação das duas abordagem é útil quando não existem dados previamente rotulados [16].

Na pesquisa descrita anteriormente foi considerada apenas duas classes de sentimento: "Positivo" e "Negativo". Noutra análise [17], testou-se a hipótese de combinação das duas abordagens considerando três classes de sentimento: "Positivo", "Negativo" e "Neutro" [17]. Foi detetado que a maioria dos dados considerados como "Neutro" eram na verdade opinativos ("Positivo" ou "Negativo"). Esta pesquisa concluiu que é necessária uma pré-avaliação dos dados classificados como neutros, antes da realização da classificação através de métodos de *Machine Learning* [17].

Nesta dissertação será avaliada a eficácia de classificação das duas abordagens de forma combinada (abordagem híbrida) de forma a testar a hipótese relatada nas últimas pesquisas descritas, que referem melhores resultados coma a utilização da abordagem híbrida. Desta forma, é aplicada a abordagem de *Lexicon-based* para a classificação do sentimento do conjunto de dados de treino utilizados nos métodos de *Machine Learning*.

2.3 Aplicação de Análise de Sentimento a tweets

A análise de sentimento em *tweets* tem sido amplamente aplicada, por exemplo, em cenário de pandemia, como no caso do vírus SARS-CoV-2. A análise de sentimento de M. T. J. Ansari e N. A. Khan [18] focou-se nos *tweets* relacionados com o termo "COVID-19". Foi possível detetar que os principais assuntos comentados sobre a vacinação COVID-19 são sobre se a vacinação ajuda a proteger vidas e se é medicamente benéfica e segura [18].

Para além de situações pandémicas, o sentimento dos *tweets* também pode ser analisado em outras situações, como é caso do sistema desenvolvido por S. Asur e B. A. Huberman [19] que utilizaram a análise de sentimento de um conjunto de *tweets* para prever as receitas de bilheteria de filmes. Foram recolhidos *tweets* relacionados a 24 filmes diferentes. Os autores concluíram que o sucesso da receita de bilheteria de um filme estava diretamente relacionado à quantidade de *tweets* associados ao filme [19]. Os autores obtiveram resultados superiores em precisão aos da Bolsa de Valores de *Hollywood*. Para esta análise foi construído um modelo de regressão linear, seguindo uma abordagem *Machine Learning* [19].

De facto, em relação a análise de sentimento de *tweets* concluiu-se que *Machine Learning* é a abordagem dominante [20]. Numa pesquisa realizada foi construído um classificador de sentimento para classificar os *tweets* como "Positivo", "Negativo" ou "Neutro".

O classificador é baseado no classificador multinominal *Naive Bayes*. Foi obtido uma precisão acima dos 81% [20].

A abordagem de *Machine Learning* necessita que um conjunto de *tweets* sejam rotulados previamente com o respetivo sentimento (“Positivo”, “Negativo” ou “Neutro”)

A rotulação dos dados de treino pode ser feita de forma manual ou como mencionado anteriormente através da utilização da abordagem *Lexicon-based*. Contudo, existem ferramentas especificas que detetam o sentimento de cada *tweet* quase em tempo real: deteção de sentimento quase em tempo real para *tweets*: *Twendz, Twitter Sentiment, TweetFeel, Twend1, Twitter Sentiment2 e TweetFeel3*. Estas ferramentas foram utilizadas em diversas pesquisas como a de L. Barbosa e J. Feng [21]. Esta pesquisa utilizou *Machine Learning* e foi proposto um método de classificação de análise de sentimento em 2 etapas para o Twitter, em que primeiro os *tweets* são classificados como subjetivos ou objetivos e posteriormente os *tweets* subjetivos são distinguidos como “Positivo” ou “Negativo” [21].

Estes sistemas descritos recorreram a *tweets* históricos para realizar as suas análises, mas, também é possível o processamento de dados em tempo real [22]. No artigo dos autores H. Wang, D. Can, A. Kazemzadeh, F. Bar, e S. Narayanan, [22] é descrito um sistema baseado em *Naive Bayes*, para análise em tempo real do sentimento público em relação aos candidatos presidenciais na eleição de 2012 nos Estados Unidos da América (EUA) [22].

Nesta dissertação serão analisados apenas *tweets* históricos. Existem diversas áreas para as quais a análise de sentimento é um apoio importante, como apresentado anteriormente através do sistema de análise de sentimento de *tweets* sobre vacinação contra a COVID-19. Mas, a saúde não é a única área de aplicação deste tipo de análise, pode ser, por exemplo, aplicado a *tweets* sobre revisões de filmes. Nesta dissertação a análise de sentimento de *tweets* será aplicada à área do mercado financeiro, mais concretamente, serão recolhidos *tweets* relacionados com o Índice de Mercado *Standard & Poor's* 500 (S&P 500).

2.4 Utilização do Twitter para previsões no Mercado Financeiro

A análise de sentimento dos *tweets*, também pode ser aplicada ao mercado financeiro, nomeadamente na previsão do seu possível comportamento [23]. Numa pesquisa o foco principal foi a previsão de indicadores de mercado financeiro como *Dow Jones Industrial Average* (DIJA), NASDAQ e S&P 500 recorrendo as mensagens do Twitter [23].

Primeiramente, foram recolhidos *tweets* durante seis meses e obtendo um centésimo do volume total de todos os *tweets*. De seguida, os *tweets* foram rotulados com sentimento

"Positivo" ou "Negativo". Após teste de três hipóteses diferentes foi concluído que os utilizadores publicam mais *tweet*s quando se sentem positivos e que, quando há alguma incerteza no mercado de ações, existe tendência em expressar sentimentos como medo ou esperança. Foi possível, ainda, concluir que quando os utilizadores expressam medo e preocupação nos seus *tweets*, no dia seguinte Índices como DIJA, NASDAQ e S&P 500 sofrem uma quebra [23].

Noutra pesquisa [24], foi utilizada a ferramenta *OpinionFinder* [16], qua classifica cada *tweet* como "Positivo" ou "Negativo", para verificar se existe correlação entre os valores do Índice de Mercado *Dow Jones Industrial Average* (DIJA) [24]. Foram analisadas 7 dimensões de humor e verificou-se que apenas algumas se relacionam com o DJIA, em mais detalhe concluiu-se que a dimensão "Calma" é um indicador DJIA. Ocorreu um grande aumento da capacidade de previsão, atingindo uma taxa de sucesso de 86,7% [24]. Nesta pesquisa concluiu-se que a análise do sentimento dos *tweets*, oferece uma adição automática, rápida, gratuita e em grande escala como ferramenta de complemento ao sistema de previsões dos Índices de mercado financeiro [24].

Baseando-se na pesquisa descrita anteriormente que obteve uma previsão de cerca de 87% [24], os autores A. Mittal e A. Goel [25] aplicaram a abordagem de *Machine Learning* a um conjunto de *tweets* relacionados com indicador de mercado financeiro DIJA para prever a correlação entre o humor dos utilizadores e os valores do Índice DJIA dos dias anteriores [25]. Como resultado foi obtido 75,56% de precisão para *tweets* e valores do Índice de mercado financeiro DJIA entre junho de 2009 a dezembro de 2009 [25]. Para além da dimensão "Calma" detetada na pesquisa descrita anteriormente [24], também se concluiu que a dimensão felicidade também é um indicador DIJA.

Como observado nos sistemas descritos anteriormente para análise de sentimento de *tweets* relacionadas com o mercado financeiro é geralmente utilizada a abordagem *Machine Learning*.

Os autores K. Xu, Y. Pang, e J. Han [26] investigaram a relação de correlação entre o sentimento de *tweets* e os retornos dos principais mercados de ações globais com base no método *Multifractal Detrended Cross-Correlation Analysis* (MF-DCCA) [26].

A análise foi realizada através da divisão entre países desenvolvidos e não desenvolvidos. Não foi encontrada relação entre o retorno do mercado financeiro e o resultado da análise de sentimento para a Coreia do Sul e no Reino Unido [26]. Contudo, também foi possível detetar uma relação de correlação entre o mercado financeiro e o resultado da análise de sentimento aplicada a *tweets* em alguns países desenvolvidos e todos os países em

desenvolvimento considerados. Esta relação de correlação é consistente ao longo do tempo, nos países desenvolvidos, no entanto, para os países em desenvolvimento foi detetado que ao longo do tempo ocorrem drásticas flutuações. Os resultados produzidos confirmam a dependência entre o sentimento e os mercados financeiros globais [26].

Nesta dissertação serão analisados um conjunto de *tweets* relacionados com a empresa *Apple* representada pela *cashtag* "$AAPL" do Índice de Mercado S&P 500, sendo analisada a correlação entre o resultado da análise de sentimento aplicada a esses *tweets* e a respetiva cotação da ação para o intervalo de datas de criação desse conjunto de *tweets*. Esta correlação será a base para a avaliação dos resultados que determinam os parâmetros que definem um utilizador *influencer*.

2.5 *Influencers* no Twitter

Com a evolução e aumento do número de redes sociais o conceito *influencers* é cada vez mais abordado. Um individuo ou entidade pode ser *influencer* de uma ou mais redes sociais. Nesta dissertação, a rede social de principal foco é o Twitter, por isso este capítulo será centrado nos *influencers* da rede social Twitter. A influência na rede social Twitter é definida como a potencialidade de uma ação de utilizador provocar uma ação por parte de outro utilizador [27]. Na pesquisa de A. Leavitt, E. Burchard, D. Fisher e S. Gilbert [27] foi concluído que o número de seguidores não tem relação direta com número de *retweets* realizados, ou seja, pode existir contas com poucos utilizadores, mas cujo seus *tweets* possuem um elevado número de *retweets* por parte dos utilizadores. Ainda se conclui que um utilizador independente pode ser mais influente do que uma entidade popularmente conhecida [27].

Noutra pesquisa foi analisado a influência dos utilizadores do Twitter com base em três medidas: o número de seguidores, o número de *retweets* de cada *tweet* do utilizador e menções ao utilizador (*mentions*) para verificar sua capacidade de gerar diálogos sobre ele [28]. Esta pesquisa assim como a anterior [28] concluiu que utilizadores com grande número de seguidores não possuem necessariamente um elevado número de *tweets*, aliás esta pesquisa ainda complementa assumindo que também não possuem necessariamente um elevado número de menções. Foi detetado que a influência não é adquirida de forma acidental e por norma requer um esforço do utilizador, por exemplo, a abordagem de um único tópico por *tweet*. Foi concluído que os utilizadores podem ser influenciadores de vários tópicos [28].

Os autores E. Lahuerta-Otero e R. Cordero-Gutiérrez [29], através de uma ferramenta de mineração de dados que combina a teoria dos grafos com a teoria de influência social,

investigaram *influencers* no Twitter para descobrir as características dos seus *tweets*. Foram analisados 3853 utilizadores que têm *tweets* sobre as empresas do ramo automóvel Toyota e Nissan [29]. Deste estudo concluiu-se que os *influencers* não costumam incorporar *links* nos seus *tweets*. Os *tweets* dos *influencers* geralmente tem um número de palavras reduzido e recorrem a um elevado número de *hashtags* e menções [29]. Ainda se concluiu que os *influencers* possuem um elevado número de seguidores que expressam as suas opiniões. Esta pesquisa em vez de detetar *influencers* analisou as propriedades em comum dos *tweets* dos utilizadores já identificados pela comunidade como *influencers*, permitindo perceber como os influenciadores comunicam com os seus utilizadores [29].

Nesta dissertação as características que definem um *influencer* são:

- Número de seguidores (*followers*);
- Número total de *tweets* publicados;
- Número de listas públicas que o utilizador é membro;
- Rácio entre o número de *followers* e número de *friends*;
- Tempo que o utilizador tem a conta.

Estes parâmetros foram definidos tendo por base as pesquisas indicadas anteriormente neste capítulo e também trabalhos relacionados com o tema desta dissertação [2], [3].

2.6 *Influencers* do mercado financeiro no Twitter

Sobre o tema desta dissertação existem diversas publicações. No artigo de D. M. Limpo [2] é desenvolvido um sistema para identificar e classificar utilizadores que podem influenciar uma determinada ação do Índice de mercado financeiro S&P 500. A maioria dos *influencers* detetados são entidades/empresas do mesmo setor da ação em estudo. Esta metodologia apresenta melhores resultados quando comparado ao serviço de recomendação (*Who-To-Follow*) presente no Twitter [2].

Como trabalho complementar do artigo referido na dissertação de D. Sousa [3] desenvolveu-se um sistema de mineração de opinião (*Opinion Mining*) sobre as ações do Índice de Mercado S&P 500 e de identificação das principais características de *influencers* do mercado financeiro. Para a análise do sentimento dos *tweets* realizados pelos utilizadores identificados como *influencers* foi utilizado o algoritmo *Naive Bayes*, que classificou os *tweets* em positivos, negativos. Primeiramente, construíram-se dois conjuntos distintos de dados e aplicou-se o algoritmo *Naive Bayes* tendo sido obtido uma precisão de 72,16% e de 75,45%. Para a deteção de *influencers* recorreu-se a aplicação do Algoritmo genético como forma de obtenção dos

valores ótimos para cada parâmetro que define o *influencer* [3]. Entende-se como parâmetro características como tempo da conta, número de *tweets*, número de seguidores, etc.

Os algoritmos genéticos são considerados um processo de pesquisa utilizados na computação para encontrar uma solução exata ou aproximada para problemas de otimização e pesquisa [30].

Os *influencers* detetados na dissertação de D. Sousa [3] e no artigo de D. M. Limpo [2].

Estes trabalhos são complementares, sendo que em ambos foram pré-estabelecidas o conjunto de características que definem um utilizador *influencer*, como por exemplo número de *tweets* e número de *followers*. O objetivo deste trabalho foi definir para cada característica, o intervalo de valores adequados para o utilizador *influencer*. Por exemplo, supondo que para a característica "Número de *followers*" o intervalo de valores obtido é de 500 a 1000, este resultado significa que neste âmbito, um utilizador definido como *influencer* deverá ter o seu número de seguidores dentro deste intervalo. Ambas as pesquisas [2], [3] concluíram que, para o âmbito em estudo, os *influencers* possuem conta com muito tempo, são bastante ativos no Twitter em termos de *tweets* publicados e possuem uma tendência de utilizar os *tweets* para falar com outros utilizadores.

O tema desta dissertação é semelhante ao tema da dissertação de D. Sousa [3] e do artigo de D. M. Limpo [2], contudo, existem diferenças, uma vez, que nesta dissertação será testada a abordagem híbrida (combinação da abordagem *Machine Learning* e da abordagem *Lexicon-based*) para a análise de sentimento. Em relação à classificação de *influencers* também será utilizado o Algoritmo genético para encontrar os valores mais adequados para cada característica que define um *influencer*.

2.7 Algoritmo Genético

Este algoritmo é baseado na Teoria da Evolução de Charles Darwin's (Seleção Natural) que defende que as espécies mais adaptadas ao ambiente em que se encontram são aquelas que sobrevivem, sendo os seus genes transmitidos para as gerações seguintes [31]. O Algoritmo genético foi criado por Jonh Holland [32], e combina a sobrevivência da solução mais apta com uma troca aleatória de informações (*crossover*) entre possíveis soluções para formar um algoritmo de pesquisa semelhante à pesquisa humana [32]. Para aplicação deste algoritmo é necessário a definição dos seguintes conceitos (Ver Figura 5) [32]:

- Indivíduo – Possível solução para o problema;

- População inicial – Primeiro conjunto de indivíduos ao qual é aplicado o Algoritmo genético;
- Cromossoma – Conjunto de características analisadas durante a função de fitness;
- Função de *fitness* – Avaliação do desempenho de cada individuo em relação às características em estudo;
- Seleção – Semelhante à Seleção Natural, os indivíduos com menor desempenho são eliminados, enquanto os indivíduos com maior desempenho têm maior probabilidade de sobreviver e transmitir a sua informação para a próxima geração;
- *Crossover* – Geração de novos indivíduos através do cruzamento entre indivíduos pertencentes à população;
- Mutação – Processo de alteração de alguns valores do individuo, este processo deve ser realizado com moderação;
- Gerações – Número de vezes que o Algoritmo genético é aplicado.

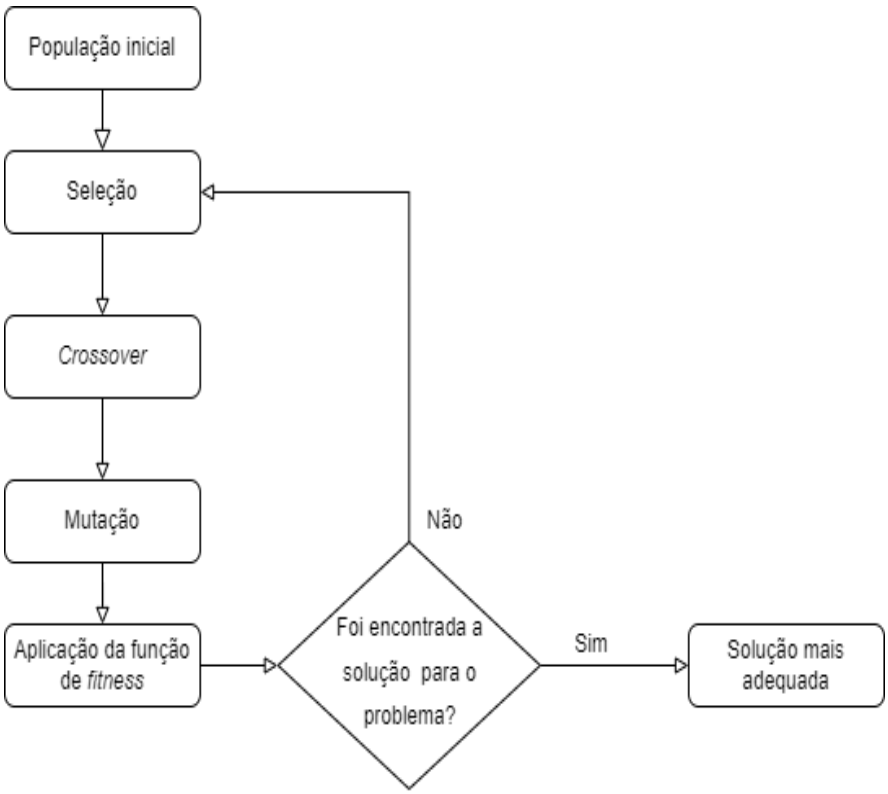


Figura 5 - Etapas do Algoritmo genético Fonte: elaboração própria

2.8 Síntese do capítulo

A partir deste capítulo é possível concluir que o Twitter é uma rede social cujas publicações consistem em mensagens, designadas de t*weets*, com número reduzido de caracteres (280 caracteres). E é esta característica que torna esta rede social uma das fontes principais de pesquisas relacionadas com análise de sentimento, sendo também, a fonte principal de dados desta dissertação.

Os *tweets* publicados pelos utilizadores expressam a opinião destes sobre determinado assunto, pode-se então dizer, que expressam um sentimento. A esta análise do conteúdo de um *tweet* é designada de análise de sentimento. Dentro desta análise destacam-se duas abordagens: *Lexicon-based* e *Machine Learning*.

Ambas as abordagens foram alvos de diversas pesquisas, tal como apresentado na Tabela 1 e deram origem a uma terceira abordagem designada de híbrida. Esta abordagem consiste na combinação de *Lexicon-based* e *Machine Learning*. Desta forma, o conjunto de treino utilizado pelos algoritmos de *Machine Learning* são rotulados através de métodos de *Lexicon-based*. Será esta a abordagem aplicada para a análise do sentimento dos *tweets* analisados nesta dissertação.

A abordagem de *Machine Learning* é constituída por vários métodos sendo que a *Suport Vector Machine* (SVM), *Neural Networks, Naive Bayes e* kNN (*k-nearest neighbors*) foram aplicadas nesta dissertação com o objetivo de testar a sua performance para a análise de sentimento do conteúdo dos tweets.

Esta análise será utilizada para determinar qual a "direção" da opinião dos utilizadores detetados com *influencers* nesta dissertação. Para esta deteção, recorre-se ao Algoritmo genético. Este algoritmo foi selecionado, uma vez que, foi implementado nos trabalhos que servem de base para esta dissertação [2], [3]. Com base num conjunto de características pré-definidas (número de *followers*, número total de *tweets* publicados, número de listas públicas que o utilizador é membro, rácio entre o número de *followers* e número de *friends* e tempo que o utilizador tem a conta) este algoritmo retorna como resultado, o intervalo de valores que para cada característica definem um utilizador *influencer*.

Para esta dissertação, *influencer* é um utilizador do Twitter cujo seus *tweets* estão relacionados com a alteração da cotação de uma determinada ação ou ações pertencentes ao Índice de Mercado S&P 500. Em resumo, pretende-se encontrar os o intervalo de valores que, para cada parâmetro, que define um *influencer* do mercado financeiro.

Durante este capítulo, foram apresentados os conceitos base para esta dissertação assim como, artigos e pesquisas relacionados com cada conceito, sendo que o resumo deste encontra-se apresentado na Tabela 1.

Tabela 1 - Resumo do capítulo 2 Fonte: elaboração própria

Conceito	**Papel nesta dissertação**	**Referências relacionadas**
Twitter	Rede social utilizada como fonte de dados.	[1],[5],[7]
Análise de sentimento	Processo utilizado para classificar os *tweets* dos utilizadores definidos como *influencers.*	[7],[9],[10],[11], [12],[13],[14],[15], [16], [17]
Abordagem Híbrida	Tipo de abordagem de análise de sentimento aplicada nesta dissertação.	[16],[17]
Influencers do mercado financeiro no Twitter	Constitui um dos objetivos desta dissertação.	[2],[3]
Algoritmo genético	Algoritmo utilizado para a deteção de *influencers*.	[30], [31],[32]

3 Metodologia

Este capítulo apresenta a descrição dos detalhes da arquitetura do sistema de classificação de *influencers* no mercado financeiro através do Twitter. Uma vez que, já existe trabalho realizado sobre este tema, a metodologia desta dissertação baseia-se na metodologia de D. Sousa [3] contudo, o módulo referente à deteção de *influencer* será baseado no artigo de D. M. Limpo [2]. Estas pesquisas tiveram como objetivo a utilização do algoritmo *Naive Bayes* e Algoritmos Genéticos para encontrar *influencers* do Twitter para prever o Índice de mercado S&P 500. A principal diferença das pesquisas citadas para o trabalho agora desenvolvido, é a aplicação nesta dissertação da abordagem híbrida de análise de sentimento.

3.1 Arquitetura do Sistema

O sistema implementado é composto por três módulos:

- Módulo de recolha e processamento de dados;
- Módulo da análise de sentimento;
- Módulo de deteção de *influencers*.

A dissertação base de D. Sousa [3] é composta pelo módulo de recolha e módulo processamento de dados, módulo de análise de dados e módulo de gestão de portfólio. Apesar do objetivo desta dissertação ser o mesmo, de forma a tornar o nome do módulo autoexplicativo nesta metodologia, o módulo de análise de sentimento corresponde ao módulo de análise de dados da dissertação base. O módulo de deteção de *influencers* deste trabalho corresponde ao módulo de gestão de portfólio da dissertação citada (Ver Tabela 2).

Tabela 2 - Comparação entre sistema da dissertação base e esta dissertação Fonte: elaboração própria

Esta dissertação	**Dissertação base [3]**
Módulo de recolha e processamento de dados	Módulo de recolha e processamento de dados
Módulo de análise de sentimento	Módulo de análise de dados
Módulo de deteção de *influencers*	Módulo de gestão de portfólio

Sendo o objetivo desta dissertação a classificação de *influencers* de mercados financeiros no Twitter, é necessário recolher as informações dos utilizadores desta rede social

desde os *tweets* realizados ao número de seguidores, entre outras informações úteis. Para analisar o nível de influência do utilizador, é necessário também ter acesso à cotação de cada ação. Pelos motivos apresentados anteriormente, o Twitter é a fonte principal desta dissertação e a fonte secundária de dados é o "*Yahoo Finance*!".

3.2 Módulo de recolha e processamento de dados

É no módulo de processamento de dados que é realizada, para um num determinado intervalo temporal, a recolha dos dados necessários:

- *Tweets*;
- Informação sobre cada utilizador do Twitter;
- Cotação de cada ação (*Adjusted Closing Price* (ACP)).

É importante realçar que após a recolha dos *tweets*, é necessário realizar um pré-processamento dos mesmos, de forma a obter apenas o conteúdo relevante para a análise de sentimento (Ver Figura 6).

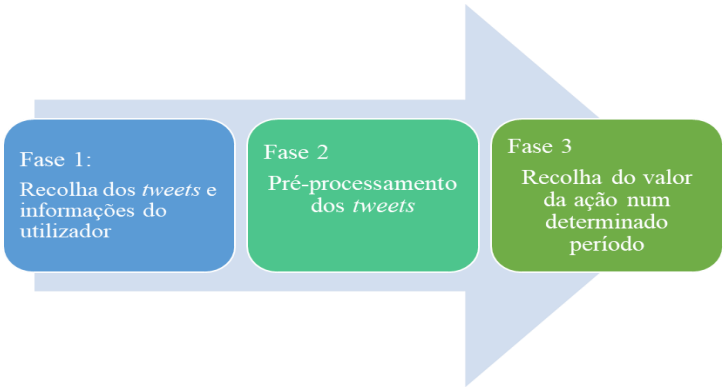


Figura 6 - Fases do módulo de recolha e processamento dos dados Fonte: elaboração própria

3.2.1 Obtenção dos *tweets*

Nesta nova abordagem metodológica, será utilizada a Linguagem de Programação *Python*, por ser uma linguagem com vários recursos ao nível de *Machine Learning* e também porque a dissertação base utilizou *Python*, o que permitiu, desde logo, validar a utilização desta linguagem de programação para a classificação de *influencers* de mercados financeiros no Twitter (Ver Tabela 3).

A dissertação base [3] recorreu ao módulo *Python Get Old Tweets* (GOT) para obter as informações necessárias da *Application Programming Interface* (API) Twitter, mas uma vez

que, este módulo foi descontinuado, nesta nova abordagem metodológica, para se poder recolher estas mesmas informações, recorreu-se ao módulo *snscrape* [3](Ver Tabela 3).

Em semelhança com o módulo GOT, o módulo *snscrape* permite recolher através da API do Twitter, *tweets* que tenham determinadas palavras-chaves. Neste caso específico, será inserido como palavra-chave os *tickers* de cada ação em análise como por exemplo "$AAPL". Para além disso, este módulo *snscrape* permite também obter *tweets* cuja data esteja dentro de um determinado intervalo, como por exemplo: "until:2021-12-06 since:2021-11-24".

O módulo referido anteriormente, tem ainda a potencialidade de extrair *tweets* de apenas um determinado utilizador, como por exemplo: "from TheTommRobinson". A limitação que se verifica no módulo GOT de restringir a recolha de *tweets* aos últimos 7 dias, não se aplica no módulo *snscrape*. Significa isto, que é possível obter todos os *tweets* de um determinado utilizador desde a criação da sua conta.

Dá-se ainda ênfase, que apenas são recolhidos os *tweets* em inglês e que contenham uma *cashtag* pertencentes às ações estudadas, por exemplo "$AAPL". Complementa-se ainda que não serão analisados *tweets* com *emojis*.

3.2.2 Pré-processamento dos *tweets*

Este tópico baseia-se por completo no pré-processamento realizado na dissertação base [3] (Ver Figura 7). O pré-processamento dos *tweets* consiste em retirar o conteúdo que não é útil para a análise de sentimento. Desta forma são retirados [3]:

- *Uniform Resource Locator* (URLs);
- Palavras com números;
- Menções a empresas (*cashtags*);
- Caracteres especiais.

Em relação as *cashtags* são retiradas uma vez que não se pretende realizar uma análise de sentimento direcionada ao nome da empresa. Contudo, os *hashtags* não são eliminados, já que, através destes os utilizadores expressam emoções, como por exemplo: #feliz. Dá-se nota, que apesar de não ser eliminado o conteúdo do *hashtag*, o # é eliminado dado ser um caracter especial. A eliminação dos caracteres especiais e de palavras com números permitem eliminar *tweets* de *spam* criados por "*web-robot*".

[3] snscrape https://github.com/JustAnotherArchivist/snscrape

A ordem de remoção dos URLs e palavras especiais não é relevante, contudo em relação aos caracteres especiais este devem ser os últimos a serem removidos. As *cashtags* incluem o caracter especial $, se os caracteres especiais forem eliminados antes das *cashtags* não é possível a deteção das mesmas.

Na dissertação base é convertido todos os *tweets* para minúsculas devido a aplicação do algoritmo *Naives Bayes*, que é *case sensitive*. Apesar de não ser o único algoritmo a ser testado nesta metodologia, também vai ser aplicado, logo, a próxima fase deste pré-processamento também inclui a conversão dos *tweets* para letras minúsculas.

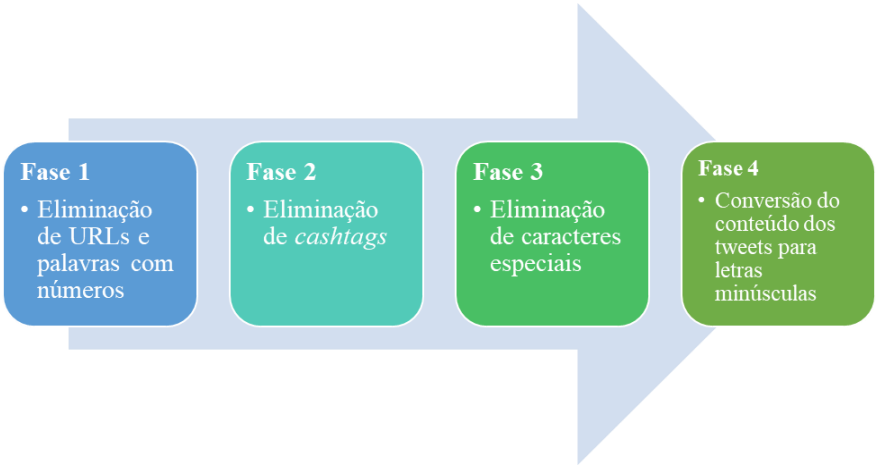


Figura 7 - Fases do pré-processamento dos tweets Fonte: elaboração própria

3.2.3 Obtenção de informação dos utilizadores

Para além do texto dos *tweets* também são recolhidas informações relacionadas com os mesmos, como o número de gostos e comentários [3]. Mas também é necessário a recolha de informações sobre o autor de cada *tweet*. Para isso, será utilizado os módulos *snscrape* e *tweepy*[4] (versão 4.8.0). Este último módulo é utilizado como complemento do primeiro de forma a mitigar as suas limitações já que com o módulo *snscrape* não é possível recolher informações mais especificas da conta do utilizador como, por exemplo, a data de criação, número de seguidores, número total de *tweets*, entre outros. Estas informações serão categorizadas em: recolhidas e calculadas. Nas primeiras, o seu conteúdo é diretamente extraído da API do Twitter através do módulo *snscrape* (Ver Tabela 3), já as segundas, são o resultado da combinação de valores extraídos [3]. Sendo o módulo de deteção de *influencers* desta dissertação baseado no artigo de D. M. Limpo [2], as informações do utilizador recolhidas nesta dissertação foram semelhantes às recolhidas por D. M. Limpo. As características consideradas não foram

[4] *tweepy* https://github.com/tweepy/tweepy

exatamente as mesmas devido às limitações de recolha dos módulos utilizados e por análise dos resultados obtidos em [2]. As informações necessárias recolhidas diretamente são:

- Número de seguidores (*followers*);
- Número total de *tweets* publicados;
- Número de listas públicas que o utilizador é membro.

Já os valores calculados são o:

- Rácio entre o número de *followers* e número de *friends*;
- Tempo que o utilizador tem a conta.

Estas características constituem o cromossoma para aplicação do Algoritmo genético sendo cada característica designada de gene (Ver Figura 8).

Figura 8 - Diferença entre o cromossoma desta dissertação vs. dissertação base [2]

3.2.4 Obtenção do valor *Adjusted Closing Price* (ACP) da ação

Para a aplicação do algoritmo de genético na deteção de *influencers*, é necessário obter informações sobre a ação em estudo. O valor de cada ação é representado pelo *Closing Price* e pelo *Adjusted Closing Price* [33]. O primeiro representa o último preço bruto transacionado antes do encerramento do mercado já o segundo, representa a cotação de fecho de uma ação tendo em contabilização as ações corporativas que podem afetar a ação, nomeadamente, desdobramento de ações e dividendos. Por isso, o ACP é uma representação mais precisa da cotação da ação [33]. Nesta dissertação será analisado o ACP da ação em estudo dentro de um intervalo de tempo de 1 mês.

Para extrair estes valores será utilizado o módulo *yfinance*[5] (Ver Tabela 3). Este pacote oferece aos desenvolvedores uma maneira de aceder informações sobre o mercado de ações [3]. A dissertação base também recorreu a este módulo e uma vez que este ainda se encontra

[5] *yfinance* https://pypi.org/project/yfinance/

disponível optou-se pela sua utilização [3]. Através dos valores de ACP é possível analisar o ganho diário da ação, que é representado pela seguinte equação (1) alicerçada na dissertação base [3] sendo também utilizada no artigo [2]. Nesta dissertação optou-se por trabalhar o resultado da função de ganho em valor decimal e não em percentagem como nas pesquisas de [2] e [3] :

$$Gain(C, d_i) = \frac{ACP(C, d_i) - ACP(C, d_{i-1})}{ACP(C, d_i)} \qquad (1)$$

Nesta fórmula C representa a ação em estudo, d_i a data que se pretende a analisar e d_{i-1} a data anterior. O ganho de uma ação para um determinado dia é obtido pela diferença entre os valores de ACP no dia em estudo e no dia anterior a dividir pelo ACP do dia anterior.

Tabela 3 - Recursos utilizados no módulo de recolha e processamento de dados: dissertação base vs. Esta dissertação Fonte: elaboração própria

	Linguagem	***Tweets***	**Dados do utilizador**	**Informações de mercado financeiro**
Dissertação base [3]	*Python*	GOT *(Get Old Tweets)*	GOT *(Get Old Tweets)*	*yfinance*
Esta Dissertação	*Python*	*snscrape*	*snscrape e tweepy*	*yfinance*

3.3 Módulo de análise de sentimento

O módulo de análise de sentimento é responsável pela análise do conteúdo de cada *tweet*. Este tópico difere da dissertação base, uma vez que, será aplicada a abordagem híbrida (*Machine Learning* e *Lexicon-based*), ao contrário da dissertação base que só recorre ao *Naive Bayes* como método de análise de sentimento (Ver Tabela 4).

Tabela 4 - Abordagens de análise de sentimento aplicadas: dissertação base vs. Esta dissertação Fonte: elaboração própria

	Abordagem	**Métodos**
Dissertação base [3]	*Machine Learning*	*Naive Bayes*
Esta Dissertação	Abordagem híbrida (Aplicação de *Lexicon-Based* para o conjunto de treino e de *Machine Learning* para o conjunto de teste)	*Navive Bayes* SVM Redes Neuronais kNN

3.3.1 Machine Learning

Para a abordagem *Machine Learning* será utilizada o *Orange* (versão 3.26.0). Esta é uma ferramenta de ciência de dados baseada no fluxo de trabalhos utilizada para a extração de dados. A abordagem *Machine Learning* será supervisionada, sendo que, para aplicação dos métodos é necessário a constituição de um conjunto de treino. O conjunto de treino consiste num conjunto de frases rotuladas como: "Positivo", "Neutro" ou "Negativo".

3.3.2 Rotulação do conjunto de treino – abordagem híbrida

Na abordagem híbrida é utilizado a abordagem *Lexicon-based* para a classificação do conjunto de treino. Através do módulo *Python* VADER é calculado o sentimento de cada *tweet* ("Positivo", "Neutro" ou "Negativo") e bem como o número de palavras positivas e negativas. Para determinar o sentimento de uma palavra, é aplicado o módulo VADER, se a polaridade da palavra for igual ou superior a 0.05 a palavra é classificada como "Positiva". Palavras com polaridade inferior ou igual a -0.05 tem sentimento negativo. Caso não sejam satisfeitas nenhuma destas condições, a palavra é considerada como "Neutra". Como referido anteriormente, também é determinado o sentimento de cada *tweet*, cujo cálculo de sentimento segue o mesmo processo.

3.3.3 Justificação para a utilização do módulo *Python* VADER

Para a classificação dos *tweets* do conjunto de treino foi utilizado o módulo VADER (*Valence Aware Dictionary for sEntiment Reasoning*) [34]. Este é utilizado para a análise de

sentimento de texto, especialmente de texto publicado nas redes socias. Para esta análise o módulo VADER baseia-se num conjunto de palavras previamente tabeladas como "Positivo", "Negativo" e "Neutro" e a partir deste conjunto prevê o sentimento de outros texto e palavras. Uma vez que, se baseia num conjunto de palavras para a análise de sentimento este módulo implementa uma abordagem *Lexicon-based*.

Existem diversos módulos *Python* baseados no léxico para análise de sentimento, contudo esta dissertação optou pela utilização do módulo VADER. A justificação para esta opção é baseada na análise realizada por C.J. Hutto e E. Gilbert, que após a análise, validação e avaliação deste módulo concluíram que este módulo apresenta uma elevada performance para a análise de sentimento de texto proveniente de redes socias [34]. Em termos de desempenho o VADER pode atingir valores de coeficiente de correlação (r = 0,881) tão elevados quanto aos valores gerados através de uma avaliação humana (r = 0,888). E, em relação à precisão de classificação o valor produzido considerando a classificação manual (F_1 = 0,84) é inferior à precisão de classificação do VADER (F_1 = 0,96), obtendo este módulo um melhor desempenho [34].

Pelos motivos referidos anteriormente e por ser um módulo de fácil acesso e compreensão, para a análise de sentimento esta dissertação selecionou o módulo VADER.

3.3.3.1 Aplicação dos métodos

Depois de rotular os *tweets*, existem três categorias de sentimento: "Positivo", "Negativo" e "Neutro". A estes *tweets* são aplicados o pré-processamento descrito anteriormente.

O resultado obtido é uma tabela cuja primeira coluna corresponde ao texto do *tweet*, a segunda coluna é o número de palavras com sentimento positivo, a terceira coluna contém o número de palavras com sentimento negativo e por fim a última coluna contem o rótulo do sentimento ("Positivo" ou "Negativo"). A tabela é posteriormente convertida para um ficheiro CSV.

Após a constituição do conjunto de treino serão aplicados os seguintes métodos de *Machine Learning* através da plataforma *Orange*:

- SVM;
- *Neural Network* (MPL)*;*
- *Naive Bayes;*
- kNN (*k-nearest neighbors*).

Sendo *Orange* uma plataforma que fornece uma interface gráfica para aplicação de métodos de *Machine Learning*, apenas é necessário realizar o carregamento do ficheiro de treino realizar a conversão para uma tabela e de seguida selecionar a variável que será o objetivo da classificação, neste caso é a variável polaridade. Por fim é realizada a análise através da matriz de confusão e do gráfico *Receiver operating characteristic* (ROC).

3.4 Módulo de deteção de *influencers*

A metodologia aplicada neste módulo segue baseia-se no sistema de D. M. Limpo [2].

Para determinar quais as características que melhor definem um *influencer* será aplicado o Algoritmo genético. Para isso, com base na função de *fitness* são selecionadas as soluções mais aptas para o problema, sendo estas designadas de pais. Cada par de pais é combinado para produzir um ou mais filhos que contem genes dos pais, gerados pela operação de *crossover*. Ainda é aplicado o processo de mutação aos filhos de modo a gerar algumas alterações nos mesmos [32].

Após esta fase, pretende-se obter a melhor solução para cada geração e para isso, os melhores pais são combinados com vista a gerar mais filhos.

O processo é repetido por um número limitado de gerações ou até uma solução satisfatória ser encontrada.

3.4.1 Representação dos indivíduos

Devido à dispersão de valores entre os utilizadores, em vez de se representar cada utilizador com um único indivíduo, este será representado por um conjunto de utilizadores.

Para cada característica analisada (gene) os utilizadores são ordenados de forma decrescente de acordo com os valores que possuem para essa característica. Por sua vez, cada conjunto é dividido em vários subconjuntos (designados de *Ranks*) [2]. Por exemplo, para a característica “número de seguidores”, o primeiro subconjunto é composto pelos utilizadores com mais seguidores e o último pelos seguidores com o menor número de seguidores [2]. Este processo repete-se para as restantes características a analisar. O número de *Ranks* deve ser entre 10 e 50. Uma vez que quanto mais utilizadores existirem em cada R*ank* maior é o intervalo de valores não é aconselhável ter apenas 10 *Ranks*. Por sua vez, também não se deve utilizar 50 *Ranks*, uma vez que, para cada *Rank* teremos um número reduzido de utilizadores reduzindo assim demasiado o intervalo de valores.

Para esta simulação, dado que, o número total de utilizadores é de 800 e de modo a não existirem nem um elevado número de utilizadores em cada *Rank*, nem um número reduzido, foram considerados 20 *Ranks* cada um composto por 40 utilizadores (Ver Figura 9).

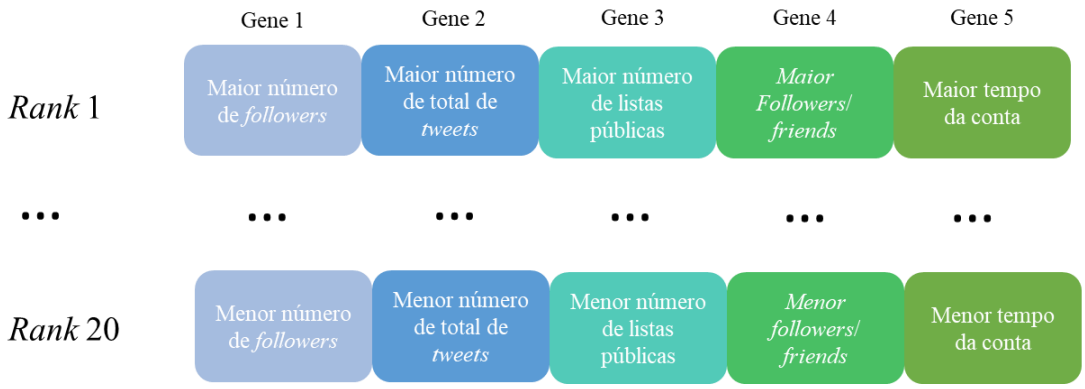


Figura 9 - Representação dos Ranks utilizados nesta dissertação Fonte: elaboração própria

Cada individuo, ou seja, cada R*ank* é composto por 5 genes (cromossoma) um para cada caraterística que se pretende analisar (Ver Figura 10).

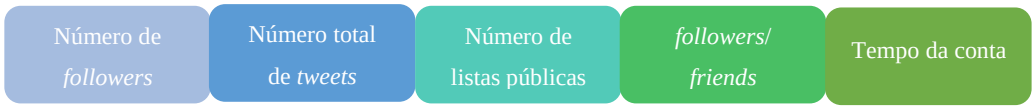


Figura 10 - Representação do cromossoma aplicado nesta dissertação Fonte: elaboração própria

3.4.2 Definição da função de avaliação: função de *fitness*

Após a definição do número de *Ranks* é necessário utilizar a função de aptidão designada de função de *fitness*.

Na função de *fitness*, representada pela equação (2), é calculado para cada utilizador pertencente ao *Rank* o impacto dos *tweets* que fez numa determinada data. Esse impacto é calculado através do produto entre a polaridade do *tweet* e soma do ganho da ACP durante os 10 dias precedentes à publicação do *tweet*. O objetivo é encontrar valores positivos, pois significa que o sentimento dos *tweets* do utilizador estava de acordo com a variação da ação. Caso contrário, se o valor for negativo, a previsão do utilizador sobre a ação não foi de acordo com a sua variação [2].

$$função_{fitness} = \sum_{i=1}^{n} (\sum_{dia=1}^{10} Gain(C, data_{tweet} + dia)) \times polaridade_{tweet\ i} \qquad (2)$$

3.4.3 Processo de seleção

Tendo por base o trabalho realizado por D. M. Limpo [2], para este processo recorre-se ao método de seleção de Tournament. Neste método são selecionados de forma aleatória *n* R*anks* para cada torneio, sendo o vencedor o indivíduo com melhor desempenho. Nesta dissertação serão considerados torneios de 3 indivíduos, para cada individuo é calculado a função de *fitness* (o valor resultante é designado de ganho) sendo o vencedor aquele com maior valor para a função (Ver Figura 11). No total são realizados 10 torneiros com 10 indivíduos vencedores, os restantes são eliminados. Os *Ranks* vencedores passam para a fase de *crossover*.

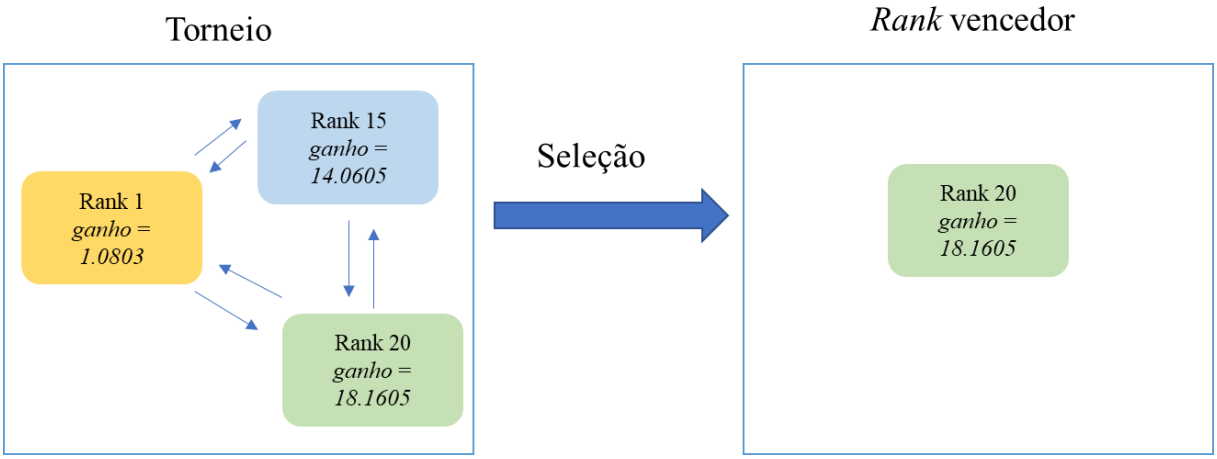


Figura 11 – Exemplo do processo de seleção desta dissertação Fonte: elaboração própria

3.4.4 Processo de *Crossover*

Após o processo de seleção dos *Ranks* vencedores são designados de pais. Na fase de crossover são gerados os filhos para cada par de *Ranks*. De forma aleatória são escolhidos 2 *Ranks* que formarão os pais, os filhos de cada par são constituídos pela junção dos genes dos seus pais. Assim como ocorreu na dissertação base [2] de D. M. Limpo, são selecionados dois pontos de corte (*Double Point Crossover*) de forma aleatória. Cada par de pais terá dois filhos.

No primeiro filho os genes entre os pontos de corte são copiados do primeiro pai, sendo os restantes genes copiados do segundo pai. Já no segundo filho é realizado o inverso, ou seja, os genes entre os pontos de corte são copiados do segundo pai (Ver Figura 12).

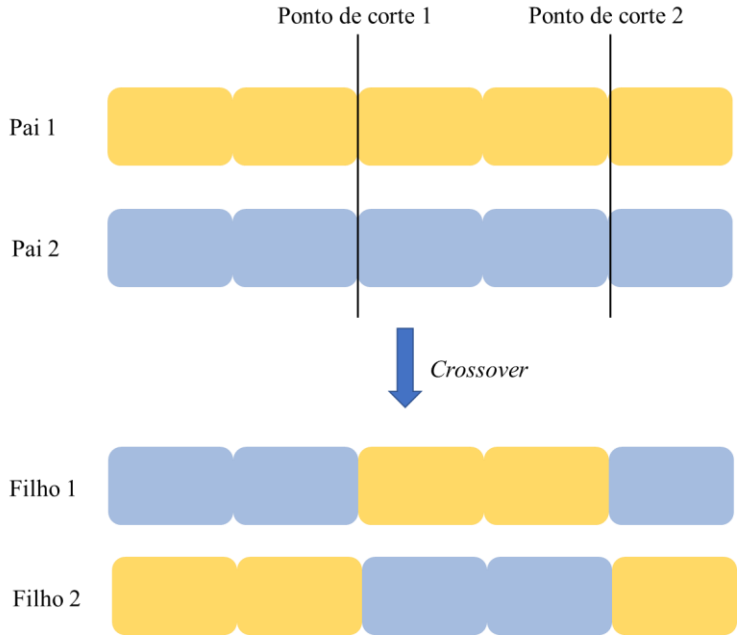


Figura 12 - Processo de Crossover desta dissertação Fonte: elaboração própria

3.4.5 Mutação

Após o processo de crossover a nova população é constituída pelo pais e filhos criados. Antes da aplicação da função *fitness* é realizado o processo de mutação. Sendo este um processo moroso é aplicada a taxa de mutação mais comumente usada com a probabilidade de $\frac{1}{5}$ [2], em que 5 representa o número de genes considerados.

3.4.6 Resumo dos parâmetros do Algoritmo genético

O Algoritmo genético aplicado nesta dissertação é composto por indivíduos que representam um conjunto de utilizadores (*Ranks*), cada individuo é composto por 1 cromossoma com 5 genes (características em análise).

A população é composta por 20 *Ranks*, sendo o número de gerações consideradas 200 (de acordo com o que foi aplicado na dissertação base [2]);

3.5 Síntese do capítulo

Neste capítulo apresentado a metodologia para a construção do sistema de classificação de *influencers* desta dissertação. Esta metodologia do sistema desta dissertação teve por base o artigo de D. M. Limpo [2] e a dissertação de D. Sousa [3]. Estes trabalhos tiveram como objetivo a classificação de *influencers* recorrendo ao algoritmo *Naive Bayes* e ao Algoritmo genético.

O sistema implementado nesta dissertação é composto por três módulos:

- Módulo de recolha e processamento de dados;
- Módulo de análise de sentimento;
- Módulo de deteção de *influencers.*

No primeiro módulo são recolhidos toda informação necessária para este sistema:

- *Tweets* (todos os *tweets* que contem a *cahstag "$AAPL"*;
- Informação sobre cada utilizador do Twitter;
- Cotação de cada ação (*Adjusted Closing Price* (ACP)).

Após esta recolha, é realizado um pré-processamento dos *tweets* recolhidos de forma a obter apenas o texto relevante para a análise de sentimento. Sendo assim, de cada *tweet* são retirados:

- *Uniforms Resource Locator* (URLs);
- Palavras com números;
- Menções a empresas (*cashtags*);
- Caracteres especiais.

Ainda neste módulo são recolhidas e calculadas para cada utilizador as características utilizadas no Algoritmo genético:

- Número de seguidores (*followers*);
- Número total de *tweets* publicados;
- Número de listas públicas que o utilizador é membro.
- Rácio entre o número de *followers* e número de *friends*;
- Tempo que o utilizador tem a conta.

No módulo de análise de sentimento através da plataforma *Orange*, são aplicados os seguintes métodos de *Machine Learning*:

- SVM;
- *Neural Network* (MPL)*;*
- *Naive Bayes;*

- kNN (*k-nearest neighbors*).

Estes métodos são de aprendizagem supervisionada, o que significa é necessário um conjunto de treino. Uma vez que nesta dissertação, recorre-se à abordagem híbrida foi utilizado o módulo *Python* VADER (abordagem *Lexicon-based*).

Por fim, no módulo de deteção de *influencers* recorre-se ao Algoritmo genético que retorna para cada característica em análise, o intervalo de valores adequado para este estudo. Desta forma, são obtidos os valores que para cada característica definem um utilizador *influencer*.

4 Validação do sistema

Neste capítulo o sistema, cuja metodologia foi definida no capítulo anterior, será testado com o objetivo de analisar e validar a sua performance. Esta simulação decorreu sobre o Índice correspondente à empresa *Apple* com *cashtag* $AAPL. Para testar o sistema foram seguidos todos os passos indicados no capítulo anterior.

4.1 Recolha e processamento dos dados

Para a aplicação da análise de sentimento, foram recolhidos os *tweets* que continham a *cashtag* $AAPL. Para esta simulação, foi considerado o intervalo de tempo de 1 mês, mais especificamente o período de 12 de janeiro de 2022 a 12 de fevereiro de 2022. Para além dos *tweets* foram recolhidas todas as informações sobre os *tweets* e utilizadores necessárias para o funcionamento do sistema.

4.1.1 Obtenção e pré-processamento de *tweets*

Para a recolha dos *tweets* recorreu-se à Twitter API através do módulo *Python* designado de *snscrape*. Considerando o período de 12 de janeiro de 2022 a 12 de fevereiro de 2022, foram recolhidos *tweets* cujo texto contém a *cashtag* $AAPL. Juntamente com o texto do *tweet* foi também recolhido o nome do utilizador e data de publicação do *tweet*. Após a obtenção deste conjunto de *tweets*, foi realizado o seu pré-processamento que consistiu na eliminação de *tweets* com URLs e/ou com palavras com números.

Este pré-processamento é realizado para eliminar o conteúdo que não contribui para a análise de sentimento. Para finalizar o pré-processamento, foram retirados deste conjunto todos os *tweets* com *emojis*, formando assim o conjunto final de *tweets*. A este foi aplicado uma conversão do seu conteúdo para letra minúscula. O conjunto final é composto com 13443 *tweets*

o que corresponde a um volume de 2454 KB. Estes *tweets* foram armazenados num ficheiro *Comma-separated values* (CSV) (Ver Tabela 5).

Tabela 5 - Número total de tweets recolhidos Fonte: elaboração própria

Número total de *tweets* recolhidos	**Volume em KB**
13 443	2454 KB

4.1.2 Obtenção de informação dos utilizadores

Após a obtenção dos *tweets*, foi criado um ficheiro CSV onde se adicionou os *usernames* dos utilizadores de cada *tweet* presente no conjunto final. Cada nome foi adicionado uma única vez não existindo *usernames* repetidos no ficheiro CSV criado. Este contém um total de 2668 *usernames*.

Com o módulo *snscrape* é possível obter os *tweets*, *usernames* e data de criação dos *tweets,* contudo para a implementação do sistema proposto nesta dissertação são necessárias mais informações sobre o utilizador. Para obter estas informações recorreu-se ao módulo *tweepy*. Com este módulo obteve-se as seguintes informações:

- Número de seguidores (*followers*);
- Número total de *tweets* publicados;
- Número de listas públicas que o utilizador é membro;
- Rácio entre o número de *followers* e número de *friends*;
- Tempo que o utilizador tem a conta.

Estas informações serão utilizadas como características para a deteção de *influencers* através de algoritmos genéticos.

O módulo *tweepy* possui limitação ao nível do número de consultas realizadas à Twitter API, por isso, para a aplicação do Algoritmo genético será aplicada apenas a um conjunto dos utilizadores recolhidos. O que significa que o número de total de utilizadores selecionados para a análise foram 800. A informação destes utilizadores ficou armazenada num ficheiro .CSV com um volume total de 39 KB (Ver Tabela 6).

Tabela 6 - Número total de utilizadores selecionados para a simulação Fonte: elaboração própria

Número total de utilizadores selecionados	**Volume em KB**
800	39 KB

4.2 Obtenção de informações de mercado

Após recolha das informações dos utilizadores é necessário analisar os valores assumidos pela ação "$AAPL" entre 12 de janeiro de 2022 e 12 de fevereiro de 2022. Para isso, através da recolha do valor de ACP, foi calculado o ganho diário da ação, representado pela equação (1), durante este intervalo de tempo. Os resultados de ACP obtidos através do módulo *yfinance* permitiram o cálculo de ganho diário da ação. Estes valores foram armazenados num ficheiro CSV, associando a cada cotação de ACP e de ganho a respetiva data.

4.3 Módulo de análise de sentimento

Neste módulo foram aplicadas as etapas necessárias para aplicação da análise de sentimento aos *tweets* recolhidos de acordo com a abordagem *Machine Learning*. Desta forma, foi necessário a geração do conjunto de teste e treino. Após esta etapa procedeu-se á rotulação do conjunto de treino através da abordagem *Lexicon-based*. Por fim, foram aplicados os métodos de *Machine Learning* selecionados.

4.3.1 Criação dos conjuntos de teste e treino

No módulo de análise de sentimento foram implementados algoritmos de *Machine Learning*. Os algoritmos selecionados pertencem a uma abordagem supervisionada, o que significa que para aprendizagem necessitam de um conjunto de *tweets* previamente rotulados ("Positivo", "Neutro" ou "Negativo") e de um conjunto de teste. Nesta simulação, o conjunto de treino é composto por 7943 *tweets* e o conjunto de teste por 5500 *tweets*.

Os *tweets* do conjunto de treino foram publicados entre 12 de janeiro de 2022 e 27 de janeiro de 2022. Já os *tweets* do conjunto de teste data de publicação pertencente ao intervalo entre 28 de janeiro de 2022 e 12 de fevereiro de 2022 (Ver Figura 13). Realça-se, então, um desfasamento temporal entre os *tweets* do conjunto de treino e os *tweets* do conjunto de teste. Este desfasamento foi aplicado uma vez que na dissertação base as datas dos *tweets* do conjunto de treino também não pertencem ao mesmo intervalo das datas dos *tweets* do conjunto de teste.

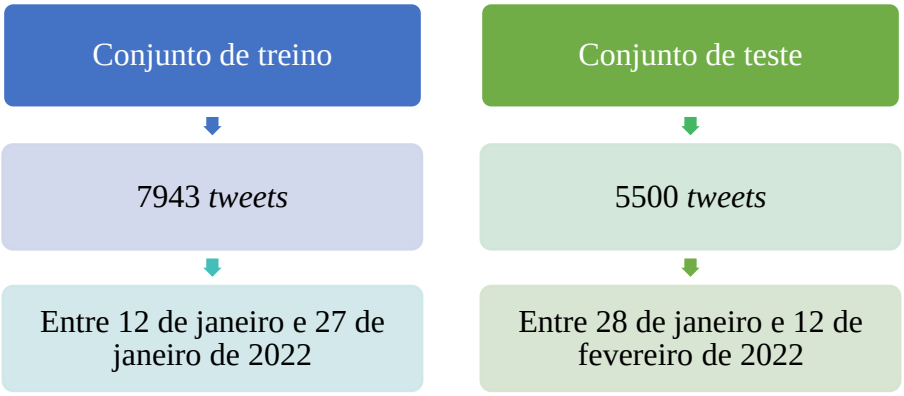


Figura 13 - Detalhes dos conjuntos de treino e de teste Fonte: elaboração própria

4.3.2 Rotulação do conjunto de treino – abordagem híbrida

Para a classificação do conjunto de treino recorreu-se ao módulo *Python* VADER. Para cada *tweet* foi calculada a polaridade de cada palavra e consequentemente foi gerado o número de palavras com sentimento "Positivo", "Negativo" ou "Neutro" presentes em cada *tweet*. Por fim, foi calculada a polaridade de todo o texto. Para um valor de polaridade superior ou igual a 0.05 a palavra/texto é classificada como "Positiva", já se for obtido um valor de polaridade inferior ou igual a -0.05 a palavra/texto é classificada com "Negativo". Caso contrário, a palavra/texto é "Neutro". Nesta simulação o número total de *tweets* rotulados como "Positivo" foi de 3926. O número de *tweets* "Negativo" foi de 1472 e por fim, 2545 *tweets* foram rotulados como "Neutro" (Ver Tabela 7).

Tabela 7 - Número de tweets rotulados como "Positivo", "Negativo" ou "Neutro" Fonte: elaboração própria

Nº de *tweets* positivos	**Nº de *tweets* neutros**	**Nº de *tweets* negativos**
3926	1472	2545

4.3.3 Aplicação dos métodos de *Machine Learning*

Para aplicação dos métodos de *Machine Learning* através da plataforma *Orange* foi realizado as seguintes fases (Ver Tabela 8):

- Fase 1: Importação do ficheiro de treino – Submissão para a plataforma do ficheiro de treino correspondente os *tweets* publicados entre 12 de janeiro e 27 de janeiro de 2022, com um total de 7943 *tweets*;
- Fase 2: Conversão do ficheiro de treino em tabela – A tabela gerada é composta por 9 colunas: nome do utilizador, texto do *tweet*, data de criação do *tweet*,

número de palavras positivas presentes no *tweet*, número de palavras negativas, total de palavras que compõem o *tweet*, valor da polaridade e rótulo de classificação (“Positivo”, “Neutro” ou “Negativo”);

- Fase 3: Importação do ficheiro de teste – Este ficheiro contém os *tweets* publicados entre 28 de janeiro e 12 de fevereiro de 2022, com um total de 5500 *tweets*;
- Fase 4: Seleção da variável objetivo: rótulo de classificação (“Positivo”, “Neutro” ou “Negativo”) é a variável objetivo;
- Fase 5: Conversão do ficheiro de treino em tabela – A tabela gerada é composta por 9 colunas: nome do utilizador, texto do *tweet*, data de criação do *tweet*, número de palavras positivas presentes no *tweet*, número de palavras negativas, total de palavras que compõem o *tweet*, valor da polaridade e rótulo de classificação (“Positivo”, “Neutro” ou “Negativo”). Na plataforma *Orange* o conjunto de teste também rotulado, sendo comparado a classificação obtida pelos métodos com a classificação “real” presente no ficheiro;
- Fase 6: Aplicação dos métodos de *Machine Learning* – Foram aplicados SVM, kNN, *Naive Bayes* e *Neural Networks;*
- Fase 7: Testagem dos resultados – Análise dos resultados de $F1$;
- Fase 8: Aplicação da matriz de confusão – Análise dos resultados e do número de classificações corretas e incorretas que cada algoritmo realizou.

Na Figura 14 é apresentado implementação dos métodos de análise de sentimento utilizados nesta dissertação, através da plataforma *Orange*. E cada componente utilizado encontra-se descrito na Tabela 8 com a respetiva indicação da fase do processo a que pertence.

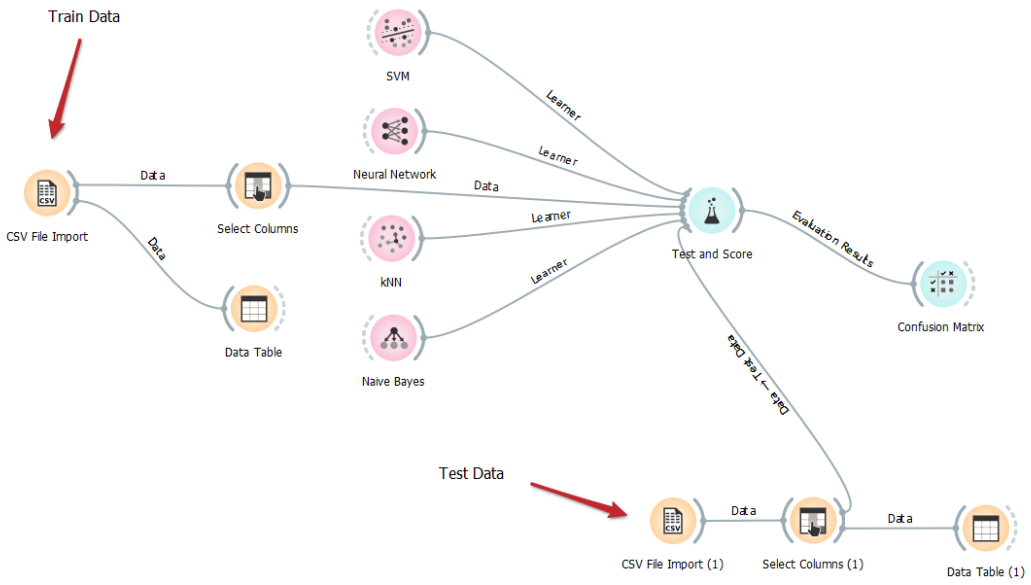
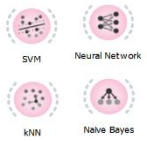


Figura 14 - Aplicação dos métodos de Machine Learning através do Orange Fonte: elaboração própria

Tabela 8 - Fases de aplicação dos métodos de Machine Learning e respetivos ícones Fonte: elaboração própria

Fase	**Função**	**Ícone *Orange***
Fase 1	Importação do ficheiro de treino	 CSV File Import
Fase 2	Conversão do ficheiro de treino em tabela	 Data Table
Fase 3	Importação do ficheiro de teste	 CSV File Import
Fase 4	Conversão do ficheiro de teste em tabela	 Data Table
Fase 5	Seleção da variável objetivo	 Select Columns

Fase 6	Aplicação dos métodos de *Machine Learning*	
Fase 7	Testagem dos resultados	
Fase 8	Aplicação da matriz de confusão	

4.3.4 Análise do resultado da análise de sentimento

Após aplicação dos métodos de análise de sentimento foi avaliado o seu desempenho através da análise do valor de *F*1 (combinação entre Precisão e *Recall*). De acordo com resultados produzidos pela plataforma *Orange* (Ver Figura 15), o algoritmo mais adequado para a classificação dos *tweets* considerados nesta dissertação é a *Neural Network*.

Model	AUC	CA	F1	Precision	Recall
kNN	0.999	0.973	0.973	0.973	0.973
SVM	1.000	0.982	0.982	0.983	0.982
Naive Bayes	0.991	0.988	0.988	0.988	0.988
Neural Network	1.000	0.999	0.999	0.999	0.999

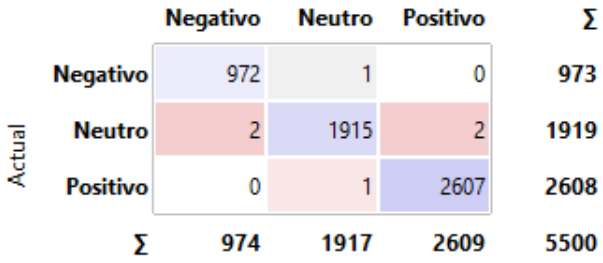
Figura 15 - Resultados dos métodos de Machine Learning Fonte: elaboração própria

Como é possível verificar, o algoritmo que tem um maior valor para o parâmetro *F*1 é a *Neural Network*, tendo sido este o algoritmo selecionado para análise de sentimento.

Analisando em detalhe a matriz de confusão produzida, verifica-se os seguintes resultados (Ver Figura 16):

- 972 *tweets* foram classificados corretamente como "Negativo", sendo que o total de *tweets* "Negativo" é de 974 *tweets;*
- 1915 *tweets* foram classificados corretamente como "Negativo", sendo que o total de *tweets* "Neutro" é de 1917 *tweets;*

- 2607 *tweets* foram classificados corretamente como "Negativo", sendo que o total de *tweets* "Positivo" é de 2609 *tweets;*



Actual	Negativo	Neutro	Positivo	Σ
Negativo	972	1	0	973
Neutro	2	1915	2	1919
Positivo	0	1	2607	2608
Σ	974	1917	2609	5500

Figura 16 - Matriz de confusão de Neural Network Fonte: elaboração própria

É de notar o elevado valor obtido para $F1$ em todos os algoritmos testados, o que se considera que seja consequência do número de *tweets* analisados. Esta questão pode ser ultrapassada com o aumento do número de *tweets* analisados. Deste modo, como trabalho futuro considera-se importante a comparar a performance destes métodos para um número mais elevado de *tweets*.

4.4 Módulo de deteção de *influencers*

Após a aplicação do Algoritmo genético para o conjunto de características definidos foi possível obter qual o *Rank* mais adequado. Ou seja, qual o intervalo de valores, para cada característica, que definem um *influencers.* Concluiu-se, que o número de seguidores de um utilizador *influencer* está entre 1842 e 3073 seguidores. Já para o número de *tweets,* estes estão entre 62698 e 1726179. Em relação às listas públicas, um utilizador *influencer* pertence a 9, 10 ou 11. Para o rácio entre o *followers* e *friends* este assume um valor entre 0,26 e 0,37. Por fim, um utilizador *influuencer* tem a conta Twitter à 3 ou 4 anos (Ver Tabela 9).

Tabela 9 - Resultados obtidos pelo Algoritmo genético Fonte: elaboração própria

Número de seguidores	**Número total de *tweets***	**Número de listas públicas**	***Followers/friends***	**Tempo de conta**
Rank 1	*Rank* 1	*Rank* 10	*Rank* 1	*Rank* 1
[31514: 556176]	[62698: 1726179]	[9:11]	[195.0: 3159.8]	[13:15]

4.5 Síntese do capítulo

Neste capítulo foi validada a metodologia do sistema desta dissertação, através da sua implementação para a deteção de *influencers* considerando o período de 12 de janeiro de 2022 a 12 de fevereiro de 2022. Por questões de limitações dos módulos de recolha utilizados, a análise focou-se nos *tweets* com a *cashtag* "$AAPL" realizados por 800 utilizadores, o que corresponde a um total de 13 443 *tweets*.

Para a recolha destes *tweets* e das informações dos utilizadores necessárias para o Algoritmo genético, recorreu-se aos módulos *tweepy* e *snscrape.*

Após a recolha destas informações foi aplicado a análise de sentimento aos *tweets* em estudo. Para isso, utilizou-se a plataforma, de *Machine Learning*, *Orange*. Uma vez que foram aplicados métodos de *Machine Learning* segundo uma aprendizagem supervisionada, foi necessário dividir os *tweets* em conjunto de treino (com 7943 *tweets*) e de teste (com 5500 *tweets*). Os *tweets* do conjunto de treino foram publicados entre 12 de janeiro de 2022 e 27 de janeiro de 2022 e os do conjunto de teste ente 28 de janeiro de 2022 e 12 de fevereiro. Os *tweets* do conjunto de teste foram previamente classificados através do módulo VADER.

Os resultados obtidos demostram elevados valores de $F1$ para todos os métodos analisados pelo que, foi selecionado as *Neural Networks* (MPL) por terem o valor de $F1$ mais elevados de todos os métodos analisados. Considera-se importante como trabalho futuro, realizar esta comparação de métodos para um número mais elevados de *tweets*.

Após esta avaliação, foi aplicado o Algoritmo genético de forma a encontrar os valores mais adequados para cada característica em estudo.

Através dos resultados obtidos pelo Algoritmo genético é possível concluir que um *influencer* é um utilizador muito ativo, ou seja, com um número elevado de *tweets* (entre 62698 e 1726179), com um número de seguidores entre 31514 e 556176. É possível ainda concluir que é um utilizador que criou a sua conta há 13, 14 ou 15 anos e que pertence a diversas listas públicas (entre 9 e 11). Por fim, pelos resultados extrai-se também que o utilizador *influencer* tem um maior número de seguidores do que de utilizadores que segue (*friends*). Estes resultados estão de acordo os trabalhos base desta dissertação [2], [3].

5 Caso de estudo

Neste capítulo serão validados os resultados obtidos no capítulo anterior através da aplicação de um exemplo prático a um dos utilizadores pertencentes ao *Rank* que constitui a solução.

Após análise dos utilizadores pertencentes a cada *Rank* que constitui a solução, foi detetado que os utilizadores jonahlupton e szaman pertencem a 4 dos 5 *Ranks* que constituem a solução. Este pertencem ao *Rank* 1 dos genes:

- Número de *followers*;
- Número total de *tweets;*
- *Followers/friends;*
- Tempo de conta.

Por isso, para estas características utilizou-se utilizadores para validação dos resultados obtidos pelo Algoritmo genético.

Em relação à característica "Número de listas públicas", o utilizador escolhido foi BP_Swing_Trader por ser o primeiro elemento do *Rank*.

A primeira etapa do caso estudo foi a análise de sentimento. Para isso recorreu-se ao módulo VADER para o conjunto de treino e à *Neural Network* para a classificação dos *tweets* em estudo.

5.1 Validação dos resultados para o *Rank* 1

Para este processo de validação foram analisados os *tweets* dos utilizadores jonahlupton e szaman. Durante o período de análise (12 de janeiro de 2022 a 12 de fevereiro de 2022), o utilizador jonahlupton realizou 22 *tweets* válidos, ou seja, sem emojis e *links*. Já o utilizador szaman realizou apenas 2 *tweets*.

Nesta dissertação, aproveitou-se esta diferença de *tweets* realizados para realizar dois tipos de análise: estratégia de curta duração (diário) e estratégia longa duração.

Na primeira, é considerado que os investidores compram e vendem no mesmo dia, tecnicamente designado de *Day trade*, onde são analisados os ACP do dia, e do dia útil seguinte. Já na segunda estratégia foi analisado o resultado do ACP no fim do mês.

Para a estratégia de curta duração foram analisados os *tweets* de szaman, já para a segunda estratégia recorreu-se aos *tweets* de jonahlupton.

5.1.1 Estratégia de curta duração

Para esta estratégia, o primeiro passo foi a recolha dos valores de ACP no dia da publicação de cada *tweet* do utilizador szaman e do dia seguinte. Após esta recolha foi determinado o sentimento de cada *tweet*.

Para o primeiro *tweet* deste utilizador, que foi publicado a 25 de janeiro de 2022, o sentimento foi "Negativo". Para este dia o ACP foi de 159.34 USD, e no dia seguinte foi de 159.25 USD. O que significa que o sentimento do *tweet* está de acordo com a alteração do valor da cotação da ação em relação ao dia seguinte à publicação do *tweet*. A título ilustrativo, supondo que um investidor comprou ações da $AAPL na abertura do mercado nesse dia (*Open Price* = 158.98 USD) e leu, nesse mesmo dia, o *tweet* deste utilizador (que foi classificado como "Negativo") e como consequência desse *tweet* vendeu as suas ações no fecho do mercado (*Close Price* = 159.7 USD), concluímos que a influência do utilizador jonahlupton foi benéfica para o investidor quando comparado com uma eventual venda dessa ação no dia seguinte, onde o ACP sofreu uma ligeira descida (159.25 USD). Com este exemplo é possível verificar que o impacto do *tweet* só se notou no ACP do dia seguinte, já que do *Open Price* ao ACP desse dia foi possível verificar um aumento de valores.

Em relação ao *tweet* de 9 de fevereiro de 2022, este também foi classificado com "Negativo". O *Open Price* desse dia foi 176.05 USD, a cotação mais elevada foi de 176.65 USD, e a cotação mais baixa foi de 174.90 USD. O ACP deste dia foi de 176.02 USD, e do dia seguinte foi de 171.87 USD. Analisando estes valores, é possível concluir que a descida do valor da cotação da ação no dia e dia seguinte ao *tweet* acompanhou a opinião do utilizador szaman.

Como exemplo do exposto, considere-se que um investidor compra ações da $AAPL quanto estas atingem a sua cotação mais baixa, ou seja, 174.90 USD. Supondo que este investidor lê o *tweet* do utilizador szaman e face ao sentimento do *tweet* decide vender as suas ações no fecho do mercado ou seja a 176.28 USD. Com esta decisão de venda, o investidor consegue um rendimento positivo com a operação. Caso o investidor não opte pela decisão de venda no próprio dia, no dia seguinte a venda não seria tão vantajosa, já que, o valor mais elevado atingido no dia seguinte foi de 175.48 USD. Dá-se nota ainda, que no dia seguinte, a cotação de fecho foi de 172.12 USD e o ACP de 171.87 USD (Ver Tabela 10).

Tabela 10 - Análise do impacto dos tweets do utilizador szaman Fonte: elaboração própria

Data do *tweet*	Sentimento	ACP do dia	ACP do dia seguinte
25/01/2022	Negativo	159.34 USD	159.25 USD
09/02/2022	Negativo	176.02 USD	171.87 USD

5.1.2 Estratégia de longa duração

O utilizador jonahlupton durante o período de 12 de janeiro de 2022 a 12 de fevereiro de 2022 realizou 22 *tweets*. Em relação ao sentimento destes *tweets*, 10 foram classificados como "Positivo", 6 como "Neutro" e 6 como "Negativo", o que significa que predominam os *tweets* com sentimento "Positivo".

Para esta estratégia, em vez de serem analisados e comparados os valores de ACP do dia do *tweet* e do dia seguinte ao *tweet*, foi realizada uma comparação entre os valores de ACP da data do primeiro *tweet* realizado durante o período analisado e os valores de ACP da data do último *tweet* realizado durante este período (Ver Tabela 11).

O primeiro *tweet* do utilizador jonahlupton foi publicado a 15 de janeiro de 2022, nesse dia o ACP foi de 169.34 USD. O último *tweet* deste utilizador foi a 5 de fevereiro de 2022 e nesse dia o ACP foi de 171.41 USD. Comparando estes dois valores é possível concluir que a opinião do utilizador jonahlupton a longo prazo estava de acordo com a cotação da ação. O mesmo se verifica para o *Open Price* e *Close Price*, tendo em consideração a comparação entre o valor de dia 15 de janeiro de 2022 e o valor de 5 de fevereiro de 2022.

Tabela 11 - Análise do impacto dos tweets do utilizador jonahlupton Fonte: elaboração própria

Data do *tweet*	Sentimento	ACP do dia	*Open Price*	*Close Price*
15/01/2022	Positivo	169.34 USD	172.54 USD	169.41 USD
18/01/2022	Negativo	169.34 USD	172.54 USD	169.41 USD
18/01/2022	Positivo	169.34 USD	172.54 USD	169.41 USD
19/01/2022	Neutro	165.77 USD	171.08 USD	165.94 USD
20/01/2022	Negativo	164.06 USD	169.68 USD	164.18 USD
20/01/2022	Negativo	164.06 USD	169.68 USD	164.18 USD
20/01/2022	Positivo	164.06 USD	169.68 USD	164.18 USD
20/01/2022	Negativo	164.06 USD	169.68 USD	164.18 USD

20/01/2022	Neutro	164.06 USD	169.68 USD	164.18 USD
20/01/2022	Neutro	164.06 USD	169.68 USD	164.18 USD
20/01/2022	Positivo	164.06 USD	169.68 USD	164.18 USD
20/01/2022	Neutro	164.06 USD	169.68 USD	164.18 USD
21/01/2022	Positivo	161.97 USD	166.33 USD	162.3 USD
21/01/2022	Negativo	161.97 USD	166.33 USD	162.3 USD
24/01/2022	Positivo	161.18 USD	162.3 USD	154.7 USD
25/01/2022	Neutro	159.34 USD	162.76 USD	157.02 USD
25/01/2022	Neutro	159.34 USD	162.76 USD	157.02 USD
27/01/2022	Negativo	158.78 USD	163.84 USD	158.28 USD
27/01/2022	Positivo	158.78 USD	163.84 USD	158.28 USD
27/01/2022	Positivo	158.78 USD	163.84 USD	158.28 USD
05/02/2022	Positivo	171.41 USD	173.95 USD	170.95 USD

Com esta análise é possível concluir que os intervalos de valores determinado para as características em análise são adequados, considerando a estratégia de longa duração.

5.2 Validação dos resultados para a característica Número de lista públicas

Para esta validação foram analisados os *tweets* do utilizador BP_Swing_Trader. Este utilizador publicou quatro *tweets* durante o período da análise. Para esta análise serão aplicadas a estratégia de curta de duração e de longa duração.

5.2.1 Estratégia de curta duração

Nesta estratégia foi analisado o impacto de cada *tweet* do utilizador na alteração da cotação da ação considerando o dia e o dia seguinte à data da publicação.

O primeiro *tweet* deste utilizador durante o intervalo da análise foi a 18 de janeiro de 2022, neste dia o ACP foi de 169.34 USD, o *Open Price* foi de 171.51 USD e o *High Price* foi de 172.54 USD. Neste dia, o utilizador publicou um *tweet* com sentimento "Negativo" e através dos valores mencionados anteriormente é possível concluir que o sentimento expresso pelo tweet do utilizador está de acordo com a alteração do valor da cotação da ação tendo em conta o dia seguinte à publicação do *tweet*. Em relação ao próprio dia o sentimento do *tweet* não de acordo com a alteração da cotação da ação uma vez que o *Open Price* foi de 171.51 USD e *High Price* foi de 172.54 USD, o que significa que tal como ocorreu na análise do R*ank* 1, o impacto do *tweet* é visível no dia seguinte à publicação do *tweet*.

Para o segundo *tweet* e para o terceiro *tweet*, publicados a 24 de janeiro de 2022 e 31 de janeiro de 2022 respetivamente, é possível observar o mesmo fenómeno, ou seja, o impacto do *tweet* do utilizador só é visível no valor da ação no dia seguinte. Ambos os *tweets* referidos anteriormente foram classificados com sentimento "Negativo".

Em relação ao último *tweet* deste utilizador, que foi publicado a 7 de fevereiro de 2022, este foi classificado como "Positivo". O ACP do dia do *tweet* foi 171.41USD e o ACP do dia seguinte à publicação do *tweet* foi de 174.57 USD, o que significa que ocorreu um aumento da cotação da ação. É possível verificar uma conformidade entre a opinião positiva expressada no *tweet* deste utilizador e a cotação da ação nesse dia, pois entre o *Open Price* e o ACP do dia registou-se uma diminuição da cotação da ação (de 160.02 USD para 161.62 USD) (Ver Tabela 12).

Tabela 12 - Análise do impacto dos tweets do utilizador BP_Swing_Trader Fonte: elaboração própria

Data do *tweet*	**Sentimento**	**ACP do dia**	**ACP do dia seguinte**	***Open Price***	***High Price***
18/01/2022	Negativo	169.34 USD	165.77 USD	171.51 USD	172.54 USD
24/01/2022	Negativo	161.18 USD	159.34 USD	160.02 USD	162.30 USD
31/01/2022	Negativo	174.30 USD	174.13 USD	170.16 USD	175.00 USD
07/02/2022	Positivo	171.41USD	174.57 USD	172.86 USD	173.95 USD

Com esta análise é possível concluir que o intervalo de valores determinado para a característica de "Número de listas públicas" é adequado, considerando a estratégia de curta duração.

5.2.1 Estratégia de longa duração

Para esta estratégia, foi realizado uma comparação entre os valores da cotação no dia do primeiro *tweet* e os valores da cotação do dia do último *tweet* realizado pelo utilizador BP_Swing_Trader.

Durante o período de analise, este utilizador realizou 4 *tweets* dos quais 3 foram classificados como "Negativo" e 1 como "Negativo". Através deste resultado é possível concluir que o utilizador esperava um aumento do valor da cotação da ação, contudo tal previsão não esteve de acordo com os resultados obtido. Uma vez que, de 18 de janeiro de 2022 para 07 de fevereiro de 2022 é possível verificar o aumento do:

- Valor do ACP do dia (de 169.34 USD para 171.41USD);

- Valor do ACP do dia seguinte (de 165.77 USD para 174.57 USD);
- *Open Price* (de 171.51 USD para 172.86 USD);
- *High Price* (de 172.54 USD para 173.95 USD).

Após análise destes resultados é possível constatar que intervalo de valores determinado a característica de "Número de listas públicas" não é o mais adequado para estratégias de longa duração.

5.3 Síntese do capítulo

Neste capítulo foram analisados os resultados obtidos pelo Algoritmo genético através da comparação dos *tweets* de utilizadores detetados como *influencers* e as alterações da cotação da ação. Foram consideradas duas estratégias: curta duração e longa duração.

Os utilizadores jonahlupton e szaman foram o foco para a validação dos valores obtidos pelo Algoritmo genético para as seguintes características:

- Número de *followers*;
- Número total de *tweets;*
- *Followers/friends;*
- Tempo de conta.

Para a estratégia de curta duração, foi possível concluir que o sentimento expresso pelos tweets do utilizador analisado (jonahlupton) estão de acordo com a alteração da cotação da ação tendo em consideração a comparação entre os valores do dia da publicação do *tweet* e do dia seguinte. Já no dia de publicação do *tweet*, é de realçar que ocorre uma discordância entre o sentimento expresso pelo tweet do utilizador e alteração da cotação da ação. Este fenômeno também ocorre para a característica "Número de listas públicas" para a estratégia de curta duração. Para esta característica o utilizador analisado foi BP_Swing_Trader.

Conclui-se que para a estratégia de curta duração os valores determinados pelo Algoritmo genético são adequados tendo em consideração a comparação entre o dia de publicação seguinte *tweet* e o dia seguinte.

Para a estratégia de longa duração, considerando a característica " Número de listas públicas", foi possível concluir os valores obtidos pelo Algoritmo genético não são os mais adequados, uma vez que a opinião expressa pelo utilizador em análise é contrária à alteração da cotação da ação. Já para as restantes características em análise, é possível concluir que os valores determinados pelo algoritmo são adequados para a estratégia de longa duração.

6 Conclusões

Nesta dissertação foi implementado um sistema de classificação de *influencers* com a finalidade principal de identificar quais as características que servem de base à definição de um utilizador do Twitter como *influencer*.

Existem diversos trabalhos relacionados sobre o tema, e esta dissertação teve como foco a otimização das técnicas já implementadas. A principal contribuição desta dissertação assenta na implementação da abordagem híbrida (*Lexicon-based* e *Machine Learning*) para análise de sentimento. De facto, conclui-se que, para o conjunto de *tweets* analisados, o algoritmo mais adequado é a *Neural Network*. As principais referências analisadas como base recorreram ao algoritmo *Naive Bayes*, e foi possível concluir que a *Neural Network* constitui uma alternativa válida para a análise de sentimento.

Em relação aos resultados obtidos, para a identificação dos parâmetros que definem um utilizador *influencer* para o mercado financeiro, é possível concluir, que estes estão de acordo com os resultados obtidos pelas pesquisas que serviram de base desta dissertação. Desta forma, é possível concluir que um *influencer* é:

- Um utilizador com elevado número de *tweets*;
- O seu número de *followers* é superior ao número de *friends*;
- Tem um elevado número de seguidores;
- Pertence a diversas listas públicas;
- Não é um utilizador recente do Twitter.

Destaca-se, também, que os resultados obtidos são mais adequados para uma estratégia de curta duração.

Com este trabalho, foi possível reforçar e otimizar os resultados obtidos pelas pesquisas anteriores relacionadas com esta temática.

7 Sugestões para trabalhos futuros

Para a deteção de *influencers* foram seguidas as principais etapas do trabalho já realizado por D. M. Limpo [2] e por D. Sousa [3], tendo sido acrescentado a utilização do módulo de *Machine Learning* da plataforma *Orange*. Com a análise de um mês de *tweets*, foi possível definir os valores adequados para cada característica de forma a definir um utilizador *influencer*.

As pesquisas realizadas e os conhecimentos adquiridos através do desenvolvimento desta dissertação podem ser ampliados futuramente através:

- Do aumento do número de características em estudo: devido às limitações atuais dos módulo *snscrape* e *tweepy* não foi possível considerar um maior número de características;
- Da inclusão como característica a analisar no Algoritmo genético, o estatuto de "verificado" de forma a perceber se esta características tem impacto nos resultados obtidos;
- Do aumento do período de análise, para a obtenção de maior número de dados;
- Do aumento do número de *tweets* analisados na análise de sentimento, de forma a comparar os resultados com os obtidos nesta dissertação.
- Da análise de sentimento e deteção de *influencers* em tempo real.

Bibliografia

[1] Z. Drus e H. Khalid, “Sentiment analysis in social media and its application: Systematic literature review” in Procedia Computer Science, 2019, vol. 161, pp. 707–714. doi: 10.1016/j.procs.2019.11.174.

[2] D. M. Limpo, “Finding Influential Twitter Users for the Stock Market Optimized by Genetic Algorithm,” 2016.

[3] D. Sousa, “Using Naïve Bayes and Genetic Algorithms to Find Influential Twitter Users to Forecast the S&P 500 Electrical and Computer Engineering Examination Committee,” 2017.

[4] N. Kanungsukkasem e T. Leelanupab, “Finding potential influences of a specific financial market in Twitter,” Proc. - 2015 7th Int. Conf. Inf. Technol. Electr. Eng. Envisioning Trend Comput. Inf. Eng. ICITEE 2015, pp. 414–419, 2015.

[5] K. Gligorić, A. Anderson, and R. West, “How constraints affect content: The case of Twitter’s switch from 140 to 280 characters,” *12th Int. AAAI Conf. Web Soc. Media, ICWSM 2018*, no. November, pp. 596–599, 2018.

[6] L. Huang e R. Bhayani, “Twitter Sentiment Analysis,” 2009. [Online]. Disponível em: https://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf. [Acedido em 10 Novembro 2022].

[7] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon-based methods for sentiment analysis,” Computational Linguistics, vol. 37, no. 2, pp. 267–307, 2011, doi: 10.1162/COLI_a_00049.

[8] A. Go, R. Bhayani, and L. Huang, “Twitter Sentiment Classification using Distant Supervision,”, vol., pp. 1–6, 2009.

[9] K. Murphy, Probabilistic Machine Learning. An introduction. MIT Press, 2022. [Online]. Disnponível em: https://probml.github.io/pml-book/. [Acedido em 16 Novembro 2022].

[10] T. Mitchell. Machine Learning. McGraw Hill, 1997.

[11] Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment Classification using Machine Learning Techniques,” EMNLP, 2002. [Online]. Disponível em: http://www.cs.cornell.edu/people/pabo/movie-review-data/. [Acedido em 16 Novembro 2022].

[12] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” Ain Shams Engineering Journal, vol. 5, no. 4, pp. 1093–1113, 2014, doi: 10.1016/j.asej.2014.04.011.

[13] sckitlearn, “Support Vector Machines”. [Online]. Disponível em: https://scikit-learn.org/stable/modules/svm.html#svm-classification. [Acedido em 10 Janeiro 2022].

[14] sckitlearn, “Neural network models (supervised)”. [Online]. Disponível em: https://scikit-learn.org/stable/modules/neural_networks_supervised.html [Acedido em 16 de Fevereiro 2022].

[15] sckitlearn, “Nearest Neighbors” [Online]. Disponível em: https://scikit-learn.org/stable/modules/neighbors.html [Acedido em 16 de Fevereiro 2022].

[16] T. Wilson, “OpinionFinder: A system for subjectivity analysis,” 2005. [Online]. Disponível em: http://nrrc.mitre.org/NRRC/publications.htm. [Acedido em 8 de janeiro 2022].

[17] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, “Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis”, 2011.

[18] M. T. J. Ansari and N. A. Khan, “Worldwide COVID-19 Vaccines Sentiment Analysis Through Twitter Content,” Electronic Journal of General Medicine, vol. 18, no. 6, p. em329, Nov. 2021, doi: 10.29333/ejgm/11316.

[19] S. Asur and B. A. Huberman, “Predicting the Future with Social Media,” Mar. 2010, doi: 10.1016/j.apenergy.2013.03.027.

[20] A. Pak and P. Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining.” [Online]. Disponível em: http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf [Acedido em 12 demarço 2022].

[21] L. Barbosa and J. Feng, “Robust Sentiment Detection on Twitter from Biased and Noisy Data,” 2010. [Online]. Disponível em: https://aclanthology.org/C10-2005.pdf [Acedido em 12 demarço 2022].

[22] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, “A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle,” Association for Computational Linguistics, 2012. [Online]. Disponível em: https://aclanthology.org/P12-3020.pdf [Acedido em 12 de março 2022].

[23] Xue Zhang, Hauke Fuehres, and Peter Gloor, “Predicting Stock Market Indicators Through Twitter ‘I hope it is not as bad as I fear’”, 2011.

[24] J. Bollen, H. Mao, and X.-J. Zeng, “Twitter mood predicts the stock market” Oct. 2010, doi: 10.1016/j.jocs.2010.12.007.

[25] A. Mittal and A. Goel, “Stock Prediction Using Twitter Sentiment Analysis.”, 2011.[Online]. Disponível em: https://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf

[26] K. Xu, Y. Pang, and J. Han, “Dynamic Cross-Correlation between Online Sentiment and Stock Market Performance: A Global View,” Discrete Dynamics in Nature and Society, vol. 2021, 2021, doi: 10.1155/2021/6674379. [Acedido em 8 d emarço 2022].

[27] A. Leavitt, E. Burchard, D. Fisher, and S. Gilbert, “The Influentials: New Approaches for Analyzing Influence on Twitter.” [Online]. Disponível em: http://www.webecologyproject.org/wp-content/uploads/2009/09/influence-report-final.pdf [Acedido em 18 de março 2022].

[28] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, “Measuring User Influence in Twitter: The Million Follower Fallacy.” [Online].Disponível em: https://ojs.aaai.org/index.php/ICWSM/article/view/14033 [Acedido em 18 de março 2022].

[29] E. Lahuerta-Otero and R. Cordero-Gutiérrez, “Looking for the perfect tweet. The use of data mining techniques to find influencers on twitter,” Computers in Human Behavior, vol. 64, pp. 575–583, Nov. 2016, doi: 10.1016/j.chb.2016.07.035.

[30] M. Kumar, M. Husian, N. Upreti, and D. Gupta, “Genetic Algorithm: Review And Application.” [Online]. Disponível em: https://ssrn.com/abstract=3529843 [Acedido em 18 de março 2022].

[31] C. Darwin, *Origin of the species.*, vol. 19, no. 34. 1859.M.

[32] D. A. Coley, “An Introduction to Genetic Algorithms for Scientists and Engineers,” An Introd. to Genet. Algorithms Sci. Eng., 1999, doi: 10.1142/3904.

[33] V. Norton, “Adjusted Closing Prices,” no. May 2011, 2011, [Online]. Disponível em: http://arxiv.org/abs/1105.2956 [Acedido em 18 de março 2022].

[34] E. Hutto, C.J. and Gilbert, “VADER: A Parsimonious Rule-based Model for,” *Eighth Int. AAAI Conf. Weblogs Soc. Media*, p. 18, 2014, [Online]. Disponível em: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewPaper/8109 [Acedido em 26 de março 2022].

Anexo 01 – Repositório GitHub

Link repositório GitHub: https://github.com/iinesalmeida/Finset-.git

Neste *link* encontra-se todo o código realizado para a implementação do sistema apresentado nesta dissertação