Universidade do Minho
Escola de Engenharia
Departamento de Informática

Ricardo Milhazes Veloso

Análise de Sentimentos para a Geração de Índices de Bem-Estar

Junho de 2022

Universidade do Minho
Escola de Engenharia
Departamento de Informática

Ricardo Milhazes Veloso

Análise de Sentimentos para a Geração de Índices de Bem-Estar

Tese de Mestrado
Mestrado Integrado em Engenharia Informática

Trabalho efetuado sob a orientação de
Orlando Manuel Oliveira Belo

Junho de 2022

AGRADECIMENTOS

Nesta pequena secção que me possibilita fazer um tributo aqueles que influenciaram positivamente o meu percuso académico e, mais concretamente, a escrita desta dissertação, gostaria de começar por agradecer à minha família. À minha mãe, pelo enorme sacrifício que fez pelos meus estudos e pelos estudos da minha irmã, colocando sempre o nosso futuro à frente do dela. Ao meu pai, por partilhar comigo o interesse nas novas tecnologias e por me dar a motivação necessária para procurar saber mais sobre esses temas. À minha irmã, pela enorme amizade que fomos desenvolvendo ao longo dos anos, especialmente durante a minha passagem na Universidade, levando-me a conhecer os Bomboémia - Grupo de Percussão da Universidade do Minho, grupo onde fiz grandes amizades e que me fez embarcar em grandes aventuras durante o meu percurso académico.

De seguida, gostaria de agradecer ao meu orientador, Professor Doutor Orlando de Oliveira Belo, pela preocupação que sempre demonstrou com a minha dissertação e com o meu percurso académico em geral. Nunca esquecerei a forma como me ajudou a definir o meu percurso na Universidade, numa altura de alguma indefinição e incerteza. Muito do que alcancei hoje deve-se a essa pequena conversa de meia hora que, apesar de curta, foi determinante. Gostaria também de agradecer ao Professor Filipe Gonçalves, atualmente meu colega na BOSCH, pela constante disponibilidade para me ajudar em tópicos relacionados com a minha vida académica e profissional e, em especial, pela sua amizade. Finalmente, um agradecimento especial ao Professor Doutor Telmo Pinto que me suportou durante uma fase mais complicada do meu percurso académico e que, ainda hoje, passados 5 anos, arranja tempo e disponibilidade para me ajudar.

Também gostaria de deixar uma palavra de agradecimento a todos os meus colegas da Bukan - Escola de Krav Magá, em especial ao meu Mestre Paulo Parente. Obrigado pela tua amizade, pelos teus ensinamentos e pelos valores que me transmitiste, valores estes que moldaram a minha personalidade e que, do meu ponto de vista, me fizeram uma pessoa melhor e mais ponderada.

Por último, resta-me agradecer a todos os meus amigos que me apoiaram e ajudaram a atingir esta meta. Foram muitos aqueles que ganhei durante estes 5/6 anos de vida académica e, por isso, estou eternamente grato. Um agradecimento especial aos mais antigos: ao Gil pela amizade única que partilho com ele e por ser o meu braço direito. Ao David, pelos momentos que partilhamos ao longo do nosso percurso no mestrado integrado em Engenharia Informática. Ao Rui, pelas noites longas nas salas 24 em busca de um milagre. E, finalmente, ao Silva, pelas aventuras no recinto do enterro e nos bares da Universidade do Minho.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity.

I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

ABSTRACT

In recent years, due to constant social and economic crises, there has been some concern regarding the quality of life and satisfaction of the population. Then, a need to create measures or criteria that would allow assessing the population's well-being, arose. Usually, these criteria are quite complex, because any one of them can be analyzed through different perspectives or dimensions, since the quality of life of a human being depends on several factors, such as health and education. The analysis of such criteria can be done through indexes. To deal with the complexity of these indexes and to create conditions that facilitate decision making, multidimensional systems can be used. These allow for a broader analysis of well-being indexes and their underlying dimensions. In this dissertation work, we will explore this area by creating a well-being analysis system based on indexes that will be calculated through the analysis of feelings expressed in texts.

KEYWORDS *Well-being Indexes, Sentiment Analysis, Text Mining, Online Analytical Processing (OLAP).*

RESUMO

Nos últimos anos, devido às constantes crises sociais e económicas, gerou-se alguma preocupação relativamente àquilo que diz respeito à qualidade de vida e da satisfação da população. Surgiu, então, a necessidade de criar medidas ou critérios que permitissem avaliar o seu bem-estar. Usualmente, estes critérios são bastante complexos, isto porque qualquer um deles pode ser analisado através de diferentes perspetivas ou dimensões, já que a qualidade de vida de um ser humano depende de vários fatores, como por exemplo a saúde e a educação. A análise de tais critérios pode ser realizada através de índices. Para lidar com a complexidade destes índices e para criar condições que facilitem a tomada de decisões, podem ser utilizados sistemas multidimensionais. Estes permitem uma análise mais ampla dos índices de bem-estar e das suas dimensões subjacentes. Neste trabalho de dissertação iremos explorar esta área através da criação de um sistema de análise de bem-estar baseado em índices que serão calculados através da análise de sentimentos expressa em textos.

PALAVRAS-CHAVE *Índices de Bem-estar, Análise de Sentimentos, Processamento de Textos, Sistemas Multidimensionais de Dados (OLAP).*

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF LISTINGS

1

INTRODUÇÃO

1.1 ENQUADRAMENTO

Hoje em dia existem vários sistemas que permitem avaliar o bem-estar da população. Um dos principais motivos para o aumento significativo destes sistemas teve origem em fevereiro de 2008, quando o 23º presidente francês, Nicholas Sarkozy, infeliz com a falta de dados estatísticos referentes ao estado atual da economia e da sociedade, pediu a Joseph Stiglitz (Presidente da Comissão), Amartya Sen (Consultora) e Jean Paul Fitoussi (Coordenador) para criarem uma comissão, mais tarde conhecida por "Comissão de Stiglitz" Stiglitz et al. (2009). Esta comissão teve como principal objetivo identificar os limites do índice GDP (*Gross Domestical Product*) como indicador de progressão económica e de progresso social. O GDP consiste num valor representativo de todos os bens e serviços produzidos no país em causa. Consequentemente, este valor foi criticado por ser um indicador fraco de bem-estar Fleurbaey (2009), Cummins et al. (2003), devido à sua desconsideração por aspetos importantes da vida das pessoas, como o respeito e a privacidade. Ainda assim, o GDP não foi totalmente dispensado por Stiglitz, que propôs criar um sistema estatístico composto por vários indicadores de bem-estar que representam uma avaliação objetiva e subjetiva da qualidade de vida da população Stiglitz et al. (2009), Iacus et al. (2015).

Desde então foram desenvolvidos imensos índices de bem-estar, todos eles com diferentes estruturas e com uma grande variedade de dimensões. Se observarmos com atenção, o que diferencia principalmente os índices de bem-estar tradicionais dos índices de bem-estar mais recentes, é a utilização de índices baseados em análises subjetivas e o aumento significativo da complexidade dos índices. Ora, com a utilização da análise subjetiva para a criação de índices de bem-estar, surgiram também algumas dúvidas no que diz respeito às fontes mais credíveis para recolher informação para análise.

Uma das formas mais simples de recolher dados que demonstrem o bem-estar de um conjunto de indivíduos, é através da realização de questionários. Vários tipos de questionários para o estudo do bem-estar subjetivo foram desenvolvidos, como, por exemplo, questionários gerais submetidos mundialmente, questionários gerais que têm um impacto localizado e questionários que apenas consideram um conjunto específico de indivíduos (ex. jovens e adolescentes) Iacus et al. (2015). Mas, o que por vezes é mais simples, nem sempre é mais eficaz. Após a realização de vários estudos, foram encontrados vários problemas relativos à utilização de questionários para medir o bem-estar, nomeadamente Iacus et al. (2015):

- a influência de uma questão ou de várias questões na qualidade das respostas;

- a frequência limitada dos questionários, que cinge o acompanhamento das flutuações nos sentimentos da população.

Ainda assim, é inteligível a importância que as opiniões das pessoas têm no cálculo de índices de bem-estar. Como sabemos, expor os nossos sentimentos nem sempre é fácil, por isso muitas pessoas recorrem à escrita para o fazerem. Assim, os sentimentos expressos em formato textual representam um veículo importante para a análise do bem-estar da população e, por essa razão, devem ser tomados em consideração.

1.2 MOTIVAÇÃO E OBJETIVOS

Desde o século XX, o mundo, tal como nós o conhecemos, foi alvo de inúmeras mudanças. Desde a industrialização até à digitalização, as tecnologias têm vindo a evoluir rapidamente, mas nem sempre privilegiando o bem-estar e a qualidade de vida da população. Segundo a Organização Mundial de Saúde Mathers and Loncar (2006), em 2030, a depressão irá representar o segundo maior problema de saúde pública. Outro indicador que revela um agravamento do bem-estar da população, está relacionado com a ansiedade crónica. Alguns estudos nessa área indicam que os níveis de ansiedade encontrados em crianças saudáveis nos anos oitenta são muito superiores aos de doentes psiquiátricos nos anos cinquenta Twenge (2000). Estes são apenas alguns indicadores relacionados com a saúde mental que expõem a fraca qualidade de vida do ser humano atualmente. Por isso, é necessário refletir e compreender que é urgente usarmos as tecnologias como contributo para uma solução para este problema, colocando o bem-estar de todos em primeiro lugar.

Os sistemas de bem-estar baseados em índices têm sido uma chave fundamental neste processo, ajudando na identificação de situações de bem-estar e de mal-estar para, posteriormente, serem tomadas decisões com o propósito de maximizar a qualidade de vidas das pessoas. Talvez a peça mais importante para a criação de um sistema de bem-estar baseado em índices, consista na recolha de dados para o cálculo destes. São precisos dados que ilustrem a qualidade de vida da população. Porém, estes são relativamente difíceis de encontrar.

Se observarmos com atenção o nosso quotidiano, a forma mais comum e intuitiva de identificarmos se uma pessoa está bem ou não é através dos sentimentos demonstrados por ela sobre um determinado assunto. Estes sentimentos podem ser expressos de formas diferentes, sendo uma dessas formas o formato textual. Para nós é relativamente simples ler um texto e identificar os sentimentos expressos sobre os vários conteúdos abordados no corpo deste. Contudo, mesmo para uma pessoa, a recolha de sentimentos, enquanto leitor, pode-se tornar complicada, já que os sentimentos expressos em texto podem ter um significado e um peso totalmente diferente para o autor. Este problema agrava-se significativamente quando a análise e a recolha de sentimentos são realizadas por um computador, já que este não possui qualquer tipo de conhecimento natural sobre linguagens. Assim, é necessário criar condições que permitam ao computador associar um texto a um ou mais sentimentos e, posteriormente, estabelecer um modelo de índices que incorpore o conhecimento extraído sobre os sentimentos associados.

Como já foi referido anteriormente, o propósito dos sistemas de bem-estar baseados em índices é apoiar a tomada de decisão tendo em conta a informação que os índices nos proporcionam. Por isso, é importante criar um sistema que analise estes índices aprofundadamente, para que seja possível tomar as decisões mais

acertadas. Tendo em conta a complexidade dos índices, nem sempre é fácil criar um sistema que cumpra os requisitos necessários.

Neste trabalho de dissertação analisou-se alguns dos modelos e técnicas existentes no domínio da análise de sentimentos e no domínio dos sistemas de bem-estar baseados em índices, com a finalidade de reduzir a complexidade destes últimos, implementando um sistema multidimensional de índices de bem-estar baseado em sentimentos, que permita analisar o bem-estar de um conjunto de indivíduos sobre um dado assunto, de acordo com as várias perspetivas de análise incorporadas no sistema.

1.3 TRABALHO REALIZADO

Neste trabalho de dissertação concebeu-se e implementou-se um mecanismo para depreender qual a opinião de um estudante relativamente a vários componentes de uma plataforma de apoio ao ensino. Análises desta natureza permitem-nos entender o grau de satisfação dos utilizadores e tomar decisões baseadas nesse grau, permitindo assim uma melhoria constante da qualidade de um determinado produto.

Usualmente, este tipo de análise é feita através de índices estáticos, em que o utilizador de uma determinada plataforma pode dar uma classificação, entre por exemplo uma estrela ou cinco estrelas, a um componente específico ou ao funcionamento geral desta. Nesta dissertação o objetivo é dinamizar esta abordagem clássica e adicionar um componente que permitisse retirar um valor entre um e cinco de um texto de opinião, garantindo assim que o espaço no qual os utilizadores podem exprimir a sua opinião livremente é valorizado.

Tendo em conta o propósito final desta dissertação, em primeiro lugar foi realizada uma pesquisa extensiva na área da análise de sentimentos para posteriormente desenvolver um método para classificar as opiniões dos utilizadores. Os métodos estudados e desenvolvidos envolvem a utilização e análise de modelos de aprendizagem automática, o processamento de conteúdo textual e a recolha de dados. De seguida, e considerando que não existem avaliações reais da plataforma, foi recolhido um pequeno conjunto de opiniões referentes à plataforma *e-learning* para testar a precisão dos modelos num cenário mais realista. Subsequentemente, a classificação destas opiniões dada pelos modelos criados foi guardada em documentos específicos, gerados automaticamente, que retratam aquilo que será uma resposta a um inquérito a que os estudantes terão que responder aquando da utilização da plataforma. Finalmente, estes valores foram apresentados num painel interativo (*dashboard*) que permite compreender a satisfação dos estudantes relativamente a, por exemplo, um tópico de estudo, em específico, ou até mesmo a satisfação de um estudante.

1.4 ESTRUTURA DA DISSERTAÇÃO

Após este capítulo introdutório, encontramos o segundo capítulo, especialmente orientado para a temática da análise de sentimentos, que descreve a pesquisa que foi realizada nesse domínio. Aqui, inicialmente, é feita uma contextualização geral acerca do tema, seguida da apresentação de um estado de arte sobre a análise de sentimentos, que incluí as diferentes abordagens utilizadas atualmente, a sua evolução nos últimos anos e, por fim, os seus desafios atuais.

Os dois capítulos seguintes explicam o processo de desenvolvimento de uma solução que permitisse calcular e visualizar os índices de bem-estar gerados. O capítulo de preparação de dados (Capítulo 3) detalha todo o processo de análise e transformação aplicado ao conjunto de dados recolhido, utilizando, maioritariamente, processos habituais em tarefas de processamento textual. Já o capítulo de classificação de sentimentos (Capítulo 4) retrata principalmente o método de cálculo dos índices de bem-estar, que, tal como a designação do próprio capítulo indica, é realizado através de algoritmos de aprendizagem automática. Para além disso, é feita uma análise dos índices gerados, considerando principalmente a taxa de acerto dos modelos desenvolvidos, bem como a sua visualização, tendo por base um sistema multidimensional de dados.

Finalmente, o capítulo de conclusões e trabalho futuro faz uma retrospetiva daquilo que foi desenvolvido ao longo desta dissertação, analisando de forma crítica os resultados obtidos e os diversos processos que foram elaborados. Além disso, também foi realizada uma reflexão sobre as melhorias que se podem aplicar aos modelos desenvolvidos bem como uma apreciação da utilidade do sistema em contextos reais, expondo assim possíveis passos para a integração do sistema no quotidiano a curto-médio prazo.

2

ANÁLISE DE SENTIMENTOS

2.1 CONTEXTUALIZAÇÃO

Nos últimos tempos, o crescimento exponencial da utilização de plataformas e de ferramentas online criou aquilo que agora denominamos de *Big Data*. Chen et al. (2014) definiram *Big Data* como um termo que representa conjuntos de dados enormes, com estruturas maiores, mais variadas e complexas. Estes conjuntos de dados são difíceis de armazenar, analisar e visualizar, mas contêm padrões e correlações (*big data analytics*) que são extremamente ricos em informação pertinente para empresas e organizações, especialmente no auxílio à tomada de decisões.

Um exemplo do aumento exponencial dos dados disponíveis atualmente, deu-se em outubro de 2012, quando, durante o primeiro debate presidencial entre o então atual presidente Barack Obama e Mitt Romney, foram publicados mais de 10 milhões de tweets no espaço de 2 horas. Entre todos estes tweets, alguns revelaram-se uma importante fonte de informação, contendo opiniões sobre diversos assuntos, como a saúde e os próprios candidatos.

Sendo assim, o tema da análise de sentimentos tem vindo a ser extensivamente explorado, representando uma solução possível para compreender as emoções, apreciações e opiniões relativamente a entidades como serviços, pessoas e produtos, expressas em textos de opinião.

Dada a natureza textual das fontes de extração de sentimentos, e a necessidade comum de aplicar técnicas de processamento de texto a estas fontes, revela-se uma relação intrínseca entre a área de processamento de linguagem natural (NLP) e a análise de sentimentos. Na realidade, a análise de sentimentos é um processo computacional pertencente à área de NLP. Sendo assim, ambas as áreas desenvolvem-se mutuamente, criando condições para contrariar desafios comuns e para se expandirem. Esta expansão revela-se promissora, sendo que, à semelhança do tópico da análise de sentimentos, a área de NLP é também de enorme interesse.

Atualmente, a análise de sentimentos é uma das áreas de pesquisa mais concorridas, com milhares de artigos escritos. Artigos estes que, frequentemente, visam a criação de novos métodos para classificar textos de opinião e que, segundo Montoyo et al. (2012), podem ser divididos, essencialmente, em 4 categorias diferentes:

- Criação de recursos para analisar sentimentos, que consiste em criar léxicos e *corpus* que expressam opiniões e anotá-los com a devida polaridade;
- Classificação de texto de acordo com a polaridade da opinião, que, usualmente se expressa em negativa, positiva ou neutra. Pode ser feita ao nível do documento, da frase ou do aspeto;

- Extração da opinião, que tem como objetivo identificar partes do texto que contenham opiniões, identificar a polaridade do sentimento expresso e determinar a entidade a que a opinião se refere;
- Análise das diferentes aplicações da análise de sentimentos, que procura identificar métodos de análise de sentimentos com elevada performance para todas as aplicações em causa, de forma a evitar a propagação de erros quando aplicados a qualquer tipo de contexto.

2.2 DEFINIÇÃO E APLICABILIDADE

A análise de sentimentos representa o estudo computacional das opiniões, atitudes ou emoções de um indivíduo relativamente a um determinado assunto, que pode representar um indivíduo ou conjunto de indivíduos, um evento ou um tópico Medhat et al. (2014). Existem 3 tipos de análise de sentimentos que se verificam, nomeadamente, ao nível do documento, da frase ou do aspeto.

2.2.1 *Nível do Documento*

A análise de sentimentos ao nível do documento é a mais simples, isto porque assume a existência de uma opinião geral, normalmente positiva ou negativa, que é expressa pelo autor do documento, sobre um assunto em específico Feldman (2013).

Existem inúmeras aplicações da análise de sentimentos ao nível do documento. Estas aplicações têm por base, normalmente, abordagens supervisionadas, dada a natureza binária da classificação (positiva ou negativa) Pang et al. (2002). Também existem algumas aplicações que usam abordagens não supervisionadas, sendo que a maior parte tem como objetivo calcular a orientação semântica de frases relevantes para a análise de sentimentos que estejam inseridas dentro do documento Turney (2002).

Ainda assim, como sabemos, um documento pode conter várias opiniões sobre o mesmo assunto. Para conseguirmos obter uma granularidade maior no que diz respeito à observação dos diferentes sentimentos expressos num documento de opinião, é necessário recorrer à análise de sentimentos ao nível da frase.

2.2.2 *Nível da Frase*

Ao nível da frase, a esmagadora maioria das implementações de análise de sentimentos utilizam apenas frases subjetivas, já que a recolha de sentimentos de frases objetivas é de extrema complexidade. Por essa razão, técnicas que permitem distinguir frases de opinião, de frases que contêm declarações factuais, são frequentemente utilizadas na fase de pré-processamento dos dados. Yu and Hatzivassiloglou (2003) criaram alguns métodos para distinguir frases subjetivas de frases objetivas. Inicialmente recorreram a uma abordagem que, partindo do pressuposto que o assunto é o mesmo, determina que uma frase subjetiva será mais similar a frases subjetivas do que a frases objetivas. Para atribuir uma classe a cada frase, usaram um sistema -

SIMFINDER Hatzivassiloglou et al. (2001) - que mede a semelhança entre duas frases através de palavras iguais, frases ou grupos de sinónimos, que expressem o mesmo conceito.

No que diz respeito à classificação de sentimentos ao nível da frase, à semelhança das implementações para classificar sentimentos ao nível do documento, as abordagens são baseadas em modelos supervisionados ou não supervisionados. Kim and Hovy (2007) classificaram mensagens relativas a opiniões relacionadas com eleições passadas através de métodos supervisionados, para a criação de um sistema que permita prever o resultado das eleições seguintes. Os dados de treino utilizados para os classificadores foram frases generalizadas, extraídas das mensagens referidas anteriormente. As mensagens são posteriormente classificadas, através da soma do valor da polaridade de cada frase presente nessa mensagem.

Por fim, é importante referir que a utilização deste nível de classificação não é muito diferente do nível do documento, especialmente se o documento representar uma simples frase.

2.2.3 *Nível do Aspeto*

Ambos os níveis anteriores de análise de sentimentos são adequados para frases ou documentos de opinião que se refiram a apenas um atributo geral de uma entidade. Contudo, muitas destas entidades têm vários atributos e, por vezes, é importante observar meticulosamente toda a informação contida nas opiniões. Surge, assim, a análise de sentimentos ao nível do aspeto.

A análise de sentimentos ao nível do aspeto tem como objetivo reconhecer as expressões de sentimento contidas numa opinião e todos os aspetos a que estas se referem. Feldman (2013)

Pontiki et al. (2016) identificaram três tarefas importantes para analisar sentimentos ao nível do aspeto: extração da entidade da opinião (1), extração dos aspetos (2) e classificação da polaridade dos aspetos (3). O exemplo seguinte representa, de uma forma geral, o processo para a análise de sentimentos ao nível do aspeto

- “Este computador tem uma placa gráfica muito boa”
  - “computador” é representativo da entidade da opinião;
  - “placa gráfica” é um aspeto da entidade “computador”;
  - a polaridade do sentimento expresso em relação ao aspeto “placa gráfica” é positiva, já que esta é adjetivada como “muito boa".

Como já foi mencionado anteriormente, uma frase de opinião pode conter várias opiniões sobre diferentes aspetos de uma determinada entidade. Para que se possa realmente identificar qual é o sentimento expresso por uma pessoa relativamente a uma entidade, na sua generalidade ou para cada um dos seus aspetos, é necessário primeiro identificar qual é a entidade a que a opinião se refere, e os seus respetivos aspetos ou atributos.

Extração da entidade da opinião (1) / Extração dos aspetos (2)

O objetivo destas duas tarefas é, então, identificar as entidades e os respetivos aspetos presentes em cada frase. Sendo assim, foram desenvolvidas algumas técnicas para proceder à extração, tanto da entidade da opinião, como dos seus aspetos, como por exemplo:

- POS (*Part-of-Speech Tagging*) Kumawat and Jain (2015) – consiste em extrair substantivos e frases nominais próximos de palavras de opinião, tendo em conta que a presença de uma palavra de opinião sugere, frequentemente, a existência de um aspeto na sua proximidade.
- Modelos Supervisionados Stanovsky et al. (2018) – também é possível abordar o problema da extração de entidades e dos seus respetivos aspetos como um problema de extração de informação. Neste tipo de abordagem são necessários dados de treino para suportar os modelos.

As estratégias referidas anteriormente são aplicadas a textos de opinião onde os aspetos são explicitamente mencionados. Isto pode nem sempre acontecer, o que torna este processo mais complexo, sendo necessária a identificação de aspetos implícitos. Para extrair aspetos implícitos, Hai et al. (2011) recorreram a regras de associação para equiparar expressões de sentimento (aspetos implícitos) a aspetos explícitos. Veja-se, por exemplo, que na frase "Este quadro é pesado", o aspeto presente é o peso, mas não está mencionado explicitamente.

Classificação da polaridade dos aspetos (3)

Após a identificação da entidade da opinião, e dos seus respetivos aspetos, é necessário compreender qual é o sentimento expresso pelo autor da opinião relativamente a esses aspetos. Este tipo de análise permite identificar, por exemplo, se a opinião de uma pessoa relativamente a uma identidade é, na sua generalidade, boa, já que para diferentes aspetos, a polaridade da opinião pode ser inversa.

Tem como objetivo classificar os sentimentos expressos sobre um determinado aspeto como positivos, negativos ou neutros. Nas secções seguintes serão apresentadas as diferentes abordagens para a classificação de sentimentos, que são úteis para esta tarefa.

2.2.4 *Áreas Aplicacionais*

Após cuidadosa análise dos diferentes níveis de aplicação da análise de sentimentos, é possível inferir que o tema da análise de sentimentos é extremamente abrangente. De qualquer forma, no âmbito desta dissertação, foi importante focarmo-nos principalmente na classificação de sentimentos. Esta pode ser dividida em 4 categorias Ravi and Ravi (2015): resolução de imprecisão em textos de opinião, análise de sentimentos entre e em várias línguas, classificação de sentimentos entre diferentes domínios e determinação da polaridade. Vejamos, de seguida, cada uma destas categorias.

Resolução de imprecisão em textos de opinião

A imprecisão é um problema extremamente comum em textos de opinião sendo que, frequentemente, estes são escritos de uma forma mais informal, expressando muitas vezes ironia e sarcasmo. Por exemplo, a frase "Boa ideia, mas agora volta ao mundo real" aparenta ser irónica, mas detetar estas características por si só é desafiador e, por isso, existem atualmente algumas iniciativas para a resolver. Tsur et al. (2010) criaram um modelo semi-supervisionado para classificar o nível de sarcasmo (valor entre um e cinco) de uma frase. Utilizaram opiniões acerca de livros, de onde retiraram atributos baseados em padrões e pontuação para construir

vetores de treino e teste. Os vetores de treino, já previamente classificados, são depois comparados com os vetores de teste, e a classificação é dada através de uma média ponderada do nível de sarcasmo dos k vetores de treino mais próximos.

Análise de Sentimentos entre e em vários idiomas

Cada idioma tem a sua identidade e, por isso, expressões de sentimento semelhantes têm diferentes valores sentimentais. Assim, é complicado obter um modelo universal eficaz para analisar sentimentos nos diferentes idiomas. Atualmente, existem várias abordagens que visam especificamente resolver este problema. Dessas várias abordagens evidenciam-se as seguintes:

- Baseada em léxicos – que permitem classificar uma opinião, traduzindo o léxico do idioma original para outro idioma, no qual já existam recursos que facilitem a classificação. Demirtas and Pechenizkiy (2013)
- Baseada em corpus – que permitem criar um *corpus* na linguagem original, devidamente anotada, para treinar um classificador estatístico que faculte a classificação de opiniões. Boiy and Moens (2009)

Classificação de Sentimentos entre diferentes domínios

Esta categoria consiste em classificar opiniões inseridas em diferentes domínios, como, por exemplo, opiniões sobre hotéis e opiniões sobre computadores, e necessita de, pelo menos, dois domínios: o domínio no qual o classificador vai ser treinado (domínio origem) e o domínio no qual a classificação vai ser realizada (domínio alvo). Existem duas formas principais de abordar este tema:

- Treinando o(s) classificador(es) com dados do domínio alvo e com dados do domínio origem. Tan et al. (2009) utilizaram FCE (*Frequently Co-Ocurring Entropy*) que, como o próprio nome indica, permite selecionar características generalizáveis que ocorrem tanto nos dados do domínio alvo como nos dados do domínio origem. De seguida classificaram os dados do domínio alvo através de um classificador treinado pelas características recolhidas pelo FCE;
- Treinando o(s) classificador(es) apenas com dados do domínio origem, e testar a sua eficácia nos dados do domínio alvo. Weichselbraun et al. (2013) criaram uma ontologia lexical que abrange diferentes domínios, através de dados de um determinado domínio origem já classificados. De seguida classificaram os dados do domínio alvo (produtos, hotéis e filmes) através dos valores da polaridade definidos na ontologia.

Determinação de Polaridade

É a identificação do sentimento expresso numa opinião. Esta determinação pode ser feita em, por exemplo, avaliações de produtos, fóruns e blogues, e é realizada através da utilização de técnicas baseadas em aprendizagem automática, baseadas em léxicos e híbridas. Este tema será explorado extensivamente nas secções seguintes.

2.3 O PROCESSO DE ANÁLISE DE SENTIMENTOS

A implementação de um sistema que consiga extrair e classificar sentimentos de uma forma eficiente e eficaz é uma tarefa de extrema complexidade. Na realidade não existe de momento nenhuma implementação que seja considerada universal para a tarefa de análise de sentimentos, isto porque cada implementação é composta por pequenos processos que, dependendo da tarefa em mãos, podem ser alterados para que seja possível atingir melhores resultados. Um exemplo de implementação foi dado por Medhat et al. (2014), e consiste num processo tipicamente utilizado para analisar sentimentos em opiniões referentes a produtos.

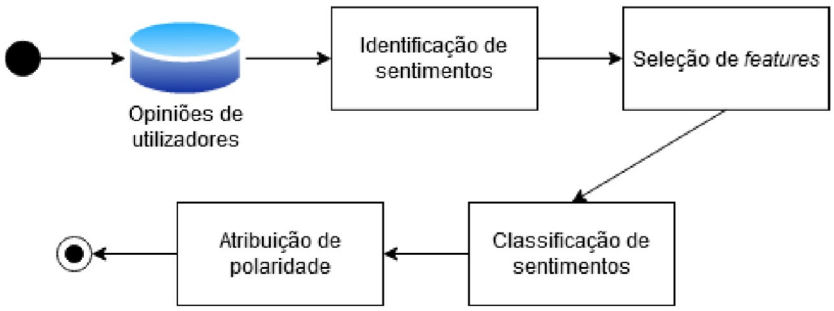

Figure 1: Processo de análise de sentimentos para opiniões referentes a produtos.

Na Figura 1 podemos ver uma versão adaptada da implementação descrita por Medhat et al. (2014), na qual encontramos as várias fases de aplicação do processo, nomeadamente, as fases de:

- Identificação de sentimentos, que permite a identificação de conteúdo subjetivo e a extração do autor da opinião. Esta tarefa é mencionada por Ravi and Ravi (2015) e permite compreender se uma frase é subjetiva ou não. Não sendo subjetiva, é improvável que a tarefa de análise de sentimentos se realize, pelo que uma frase objetiva é baseada em factos e não em opiniões. Existem raras exceções em que frases objetivas expressam sentimentos, e essas exceções vão ser abordadas mais à frente.
- Seleção de características, que possibilita um aumento na eficácia do modelo, reduzindo algum ruído que possa vir incluído nos dados originais. Esta fase pode ser incluída numa fase mais abrangente, a fase de pré-processamento dos dados.
- Classificação de sentimentos, que consiste na utilização de métodos de aprendizagem automática, baseados em dicionários ou híbridos para cotar, de uma forma quantitativa ou qualitativa, o(s) sentimento(s) expressos pelo autor de uma opinião.
- Atribuição de polaridade, que consiste em atribuir uma conotação negativa, positiva ou neutra às opiniões, através da cotação das opiniões dada pelo classificador.

O processo apresentado por Medhat et al. (2014) é um processo com já bastante conteúdo, cumprindo assim os requisitos para ser considerado um bom sistema de análise de sentimentos. A fase de classificação de

sentimentos é, na realidade, a única fase indispensável a qualquer outro processo de análise de sentimentos, sendo que permite atingir o objetivo final que é comum a todos os processos: atribuir uma cotação às opiniões. Todas as outras fases podem não fazer parte do processo final de análise de sentimentos, mas não é por isso que não deixam de ser relevantes.

Uma fase que está subentendida no processo definido por Medhat et al. (2014) é a fase de aquisição de dados. Esta é uma fase de extrema importância e, por essa razão, deve ser mencionada como parte do processo de análise de sentimentos Vinodhini and Chandrasekaran (2012). Na realidade, esta fase é indispensável, já que sem dados não é possível treinar nem testar os classificadores incluídos na fase de classificação de sentimentos. Como é óbvio, é importante treinar os classificadores com dados que contenham características textuais relevantes para a análise de sentimentos e, por isso, deve-se sempre tentar recolher dados de fontes de qualidade.

Tendo em conta as diferentes fases do processo de análise de sentimentos mencionadas nos artigos aqui referidos, foi composto o seguinte processo, processo este que se equivale ao apresentado anteriormente:

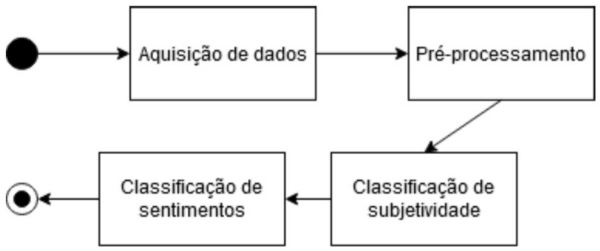

Figure 2: Processo geral de análise de sentimentos.

O processo de análise de sentimentos apresentado na Figura 2 incluí as etapas mais comumente aplicadas para determinar a polaridade dos sentimentos contidos numa opinião, etapas estas que vão ser aprofundadas na restante redação deste capítulo.

Aquisição de dados

A fase de aquisição de dados é uma fase extremamente importante para o processo de análise de sentimentos. É nesta fase que o criador do processo deve relacionar fontes existentes (as que tem acesso, obviamente) tanto quanto possível para recolher informação, com o objetivo principal do processo. Na prática, isto significa que, se o objetivo for criar um processo de análise de sentimentos para determinar a satisfação dos clientes sobre os produtos de uma empresa, é usual recolher informação em locais onde estejam disponíveis avaliações sobre os seus produtos. Para além disso é importante estabelecer qual é a complexidade pretendida para a fase de aquisição de dados, já que, por vezes, a recolha de dados é mais exigente quando as fontes são mais ricas em termos de integridade das opiniões.

Pré-processamento

Na fase de pré-processamento, o objetivo principal é criar condições para que os classificadores aplicados na fase de classificação de sentimentos tenham elevada performance. Performance esta que deve ser melhorada tanto

ao nível da rapidez dos classificadores, como ao nível da qualidade dos resultados. Existem vários passos que podem ser incluídos na fase de pré-processamento. Os mais comuns são: *tokenization*, remoção de *stopwords* e *stemming*.

Classificação de subjetividade

Como já foi referido anteriormente, compreender se as frases utilizadas para treinar os classificadores são subjetivas é uma tarefa muito importante. A subjetividade permite às pessoas expressar sentimentos e emoções, que são a fonte principal de informação para sistemas de análise de sentimentos.

Classificação de sentimentos

A fase de classificação de sentimentos é, sem dúvida, a mais importante. Permite determinar a orientação de um sentimento de um determinado texto em duas ou mais classes. Pode ser realizada através de diferentes abordagens, tais como aprendizagem automática, baseada em dicionários ou mesmo híbrida. Nas secções seguintes, todas estas fases vão ser exploradas ao detalhe, apresentado diferentes abordagens para cada uma delas.

2.3.1 *Aquisição de Dados*

A fase de aquisição de dados consiste, como o seu próprio nome indica, em recolher dados de diversas fontes para, no caso da análise de sentimentos, alimentar o sistema nas fases de classificação de subjetividade e de classificação de sentimentos. Os dados poderão servir para treino caso os modelos implementados necessitem, mas são sempre indispensáveis para testar a qualidade dos modelos utilizados.

Esta fase, como já foi indicado anteriormente, é extremamente importante para criar um bom processo de análise de sentimentos. Com dados de alta qualidade é possível criar sistemas mais eficazes e, por isso, é importante essencial quais são as fontes de dados mais indicadas. Esta identificação depende principalmente do objetivo dos modelos, mas também é importante averiguar se os dados contidos nessas fontes são íntegros. Sendo assim, estas são algumas fontes, identificadas por Vinodhini and Chandrasekaran (2012), nas quais podemos recolher opiniões:

- Blogues e micro blogues – atualmente, os blogues são utilizados com imensa frequência. Os utilizadores expressam diariamente emoções e sentimentos em blogues, criando assim uma fonte de dados vasta e também de qualidade, já que, aqui, os utilizadores se podem expressar livremente sem influências externas, sendo possível garantir um acompanhamento constante das flutuações dos sentimentos de um indivíduo ou conjunto de indivíduos. Do ponto de vista da sua utilidade, os blogues servem praticamente qualquer tipo de objetivo que um sistema de análise de sentimentos tenha, mas, tendo em conta a sua natureza mais expressiva, serão indicados, por exemplo, para a criação de indicadores de bem-estar. Kim and Hovy (2007), como já foi referido anteriormente, criaram um sistema para prever eleições. Neste contexto, é importante referir que estes recolheram os dados para treino dos modelos integrantes do sistema de um *website*. Este *website* é a página de um projeto para prever as eleições, que dá a

possibilidade aos utilizadores de darem a sua opinião sobre as mesmas, funcionando, assim, como um estilo de blogue.

- Sites de avaliações – nos últimos anos, plataformas que permitem aos consumidores avaliar os produtos que adquiriram têm vindo a crescer. É frequente os consumidores antes de comprarem um produto procurarem opiniões sobre ele para auxiliar a tomada de decisão. Estas plataformas são uma fonte de dados muito importante para as empresas isto porque, e remetendo para a secção de utilidade, permitem às empresas recolher opiniões dos consumidores sobre os seus produtos de uma forma simples, garantindo assim uma melhor compreensão do mercado em que estão inseridas, para que possam ser competitivas e capazes. Sendo assim, a sua utilidade está realmente relacionada com este tipo de objetivo empresarial. Kang et al. (2012) recolheram opiniões através de um *web-crawler* (termo que representa processos de recolha de dados da *web*), recorrendo assim a sites de avaliações para encontrar as opiniões pretendidas.

- Conjuntos de dados – por vezes não é fácil criar processos para recolher dados em sites ou blogues, já que isto implica a criação de *web-scrappers* ou *web-crawlers*, que podem ser extremamente complexos. Existem então conjuntos de dados pré-definidos que, no contexto da análise de sentimentos, já contêm opiniões de indivíduos, dando a possibilidade a quem está a construir um sistema de análise de sentimentos de não ter a necessidade de recorrer a processos de recolha de dados. Acaba por ser um método mais simples e rápido, mas nem sempre muito aconselhável, especialmente quando o sistema de análise de sentimentos está inserido num contexto específico, diferente daquele a que o conjunto de dados está associado. Basari et al. (2013) usaram opiniões sobre filmes para testar um sistema híbrido de análise de sentimentos. Estas opiniões foram recolhidas inicialmente do Twitter (blogues e micro blogues). Porém, de momento, já estão tratadas e foram agrupadas num conjunto de dados que é extremamente utilizado pela comunidade científica para testar a eficácia de sistemas de análise de sentimentos.

Após uma análise cuidadosa do processo de aquisição de dados, é possível inferir que não existe uma fonte melhor do que as outras. Cada uma deve ser aplicada consoante o objetivo do sistema de análise de sentimentos em questão. Resumidamente, um sistema que pretenda obter opiniões sobre um determinado produto deve optar por recorrer a sites de avaliações ou blogues, e um sistema que pretenda, por exemplo, apenas analisar a eficácia de determinados modelos, pode recorrer a conjuntos de dados previamente criados.

2.3.2 *Pré-Processamento*

A fase de pré-processamento é importantíssima para reduzir a complexidade das opiniões e simplificar o processo de análise de sentimentos, melhorando a qualidade dos dados iniciais. Tendo em conta que estes dados estarão em formato textual, logo são dados não estruturados, devem ser aplicadas técnicas que permitam extrair informação útil para os modelos que os vão consumir. Ora, nesta fase podem ser aplicadas diversas técnicas de processamento textual, sendo que estas devem ser selecionadas tendo sempre em grande consideração o objetivo da solução implementada. Vejamos, então, algumas dessas técnicas.

Tokenization

Tokenization consiste em partir uma frase em palavras, pequenas frases, símbolos ou outro tipo de *tokens*, para facilitar a análise individual de cada uma. Existem diferentes métodos para proceder à *tokenization* de frases. O método mais comum consiste em separar uma frase por espaços. Na realidade, nem sempre, este método é o mais indicado, já que não tem em conta palavras imediatamente seguidas de um sinal de pontuação. Sendo assim, o mais óbvio seria separar a frase não só por espaços, mas também por pontuação. Ainda assim, esta abordagem pode ser problemática porque implica a separação de indicadores importantes como emoticons. Na Tabela 1 podemos ver um exemplo de *tokenization* de uma frase.

Frase	O dia hoje está muito agradável
Token	"O", "dia", "hoje", "está", "muito", "agradável"

Table 1: Exemplo de *tokenization* de frases.

No contexto da análise de sentimentos, o processo de *tokenization* permite a análise do peso que cada *token* possui em termos de demonstração de emoções, facilitando assim a implementação de técnicas como *stemming* e remoção de *stopwords*. Porém, o facto de se analisar os *tokens* um a um, retira toda a possibilidade de analisar a frase com um todo, colocando em causa processos de, por exemplo, identificação de negação e sarcasmo.

O'Connor et al. (2010) recorreram a *tokenization* para criar um sistema de exploração de tweets por tópico. O sistema de *tokenization* é bastante robusto já que trata, não só palavras, pontuação, e abreviações, mas também caracteres importantes no contexto do Twitter como *hashtags*.

Remoção de *stopwords*

A remoção de *stopwords* consiste em remover, de textos, palavras que não são consideradas relevantes para as aplicações de mineração de textos. Esta remoção reduz imenso a dimensão de termos existentes no texto, tornando assim os textos mais claros e simples Kannan et al. (2014). Na realidade, este processo funciona extremamente bem com o processo de *tokenization*, já que este proporciona as condições necessárias para analisar com mais facilidade cada palavra contida no texto, permitindo assim remover as palavras que não são relevantes para o objetivo em causa.

Existem diferentes métodos para a remoção de *stopwords*, sendo que o principal consiste na utilização de dicionários ou listas pré-definidas compostas por *stopwords*, para depois fazer a correspondência entre as palavras do texto e as palavras contidas nas listas ou dicionários. Usualmente estas listas contêm preposições e pronomes que, de facto, não exprimem qualquer tipo de importância para qualquer tipo de aplicação de mineração de textos. A tabela 2 apresenta um exemplo de uma lista pré-definida, uma opinião não processada e um resultado proveniente da correspondência entre as palavras contidas na lista e as palavras da opinião.

Este processo é importante para a análise de sentimentos, já que reduz a complexidade das opiniões que vão alimentar os classificadores de subjetividade e de sentimentos, reduzindo assim o ruído nos dados iniciais. À semelhança do processo de *tokenization*, é impossível analisar a opinião inicial como um todo, já que muitas vezes este processo retira-lhe o sentido e o significado.

Lista	"ele", "ela", "também", "até"
Frase	Ela e ele também se dão muito bem
Token	e dão muito bem

Table 2: Exemplo remoção de *stopwords*.

Stemming

O método de *stemming* consiste em transformar uma palavra na sua palavra raiz. O objetivo principal do processo de *stemming* é melhorar a eficiência de aplicações de minerações de textos que utilizem pesquisa e indexação, através da redução em massa da dimensão dos termos Jivani et al. (2011). Esta redução normalmente não afeta a qualidade dos dados já que, frequentemente, se observa que palavras morfologicamente parecidas têm as mesmas interpretações semânticas. Contudo, é importante ter este ponto em consideração e manter separadas as palavras que não têm o mesmo significado.

Usualmente, os métodos de stemming envolvem a remoção de sufixos e prefixos das palavras para, posteriormente, lhes atribuir a sua palavra raiz. A figura 3 mostra um exemplo geral de aplicação de *stemming*, adaptado do exemplo apresentado por Kannan et al. (2014).

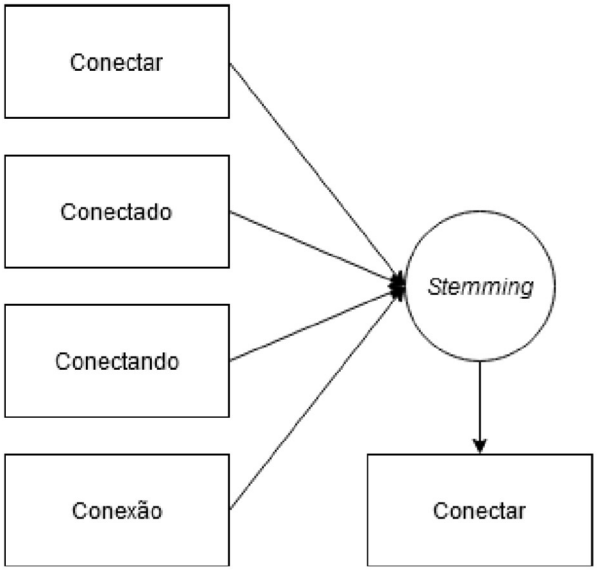

Figure 3: Ilustração geral de um processo de *stemming*.

Relativamente à análise de sentimentos, as vantagens deste processo são similares às vantagens dos dois processos anteriores. Realmente, este processo vai facilitar o trabalho aplicado pelos classificadores, já que a dimensão do espaço dos termos que vai ter que analisar é sensivelmente mais pequena.

Ainda que seja um procedimento de enorme importância e relevância, é necessário ter em consideração a possibilidade de ocorrência de *over-stemming* ou *under-stemming*. *Over-stemming* consiste em cortar palavras em demasia, o que, consequentemente, leva a duas ou mais palavras ficaram reduzidas à mesma palavra raiz,

quando na realidade as suas respetivas palavras raiz são diferentes. Já *under-stemming* representa exatamente o contrário, ou seja, reduzir duas ou mais palavras a palavras raízes diferentes, quando na realidade estas partilham a mesma palavra raiz. Uma forma de contrariar esta possibilidade é através da análise da semântica, da sintaxe e da classe de todas as palavras, aplicando apenas *stemming* às palavras mais indicadas Jivani et al. (2011).

Como é possível verificar, a maioria dos procedimentos realizados na fase de pré-processamento têm o objetivo de reduzir a dimensão do espaço dos termos, quando aplicados na área da análise de sentimentos. Por outras palavras, o objetivo principal é reduzir a complexidade dos dados mantendo a integridade das opiniões, criando assim condições para aumentar a eficácia dos modelos de classificação. Outros procedimentos que podem ser realizados nesta fase são, por exemplo, a identificação de negações, que permite analisar a alteração da orientação de uma opinião, e a identificação de palavras ou frases de opinião, como "bom" e "mau".

2.3.3 *Classificação de Subjetividade*

A classificação de subjetividade consiste em distinguir conteúdo textualmente subjetivo de conteúdo textualmente objetivo. É por isso importante definir primeiro o que é conteúdo objetivo e o que é conteúdo subjetivo. Assim, podemos afirmar que:

- Conteúdo objetivo – é composto por conteúdo factual, isto é, apresenta factos sobre uma entidade através das suas características.
- Conteúdo subjetivo – tem como características sentimentos e emoções, já que expressa uma opinião relativa a uma entidade ou às suas características.

A fase de classificação de subjetividade é realmente importante na produção de sistemas de análise de sentimentos, isto porque o conteúdo subjetivo é indicador da presença de emoções e de sentimentos. Ainda assim, é preciso ter em consideração a possibilidade da presença de sentimentos e emoções em conteúdos objetivos, algo que será explorado com mais afinco neste capítulo. Em alguns casos as opiniões podem ser mais complexas, especialmente quando abordam temas como a política. Como tal, é importante compreender quais são os tipos de subjetividade existentes, já que a classificação de subjetividade por si só, pode não ser suficiente. Maks and Vossen (2012) consideram que a subjetividade pode ser expressa de duas formas:

- Subjetividade do orador – que consiste na representação dos sentimentos, emoções e perceções do orador. Por exemplo, a frase "Ele é um mentiroso" representa uma atitude negativa do orador para com "ele";
- Subjetividade da personagem – que consiste na representação dos sentimentos, emoções e perceções da personagem. No caso da frase "O ódio dele pela religião" podemos detetar uma representação de uma atitude negativa da personagem ("dele") para com a religião, enquanto que na frase "Eles estão entusiasmados pelo serviço" identificámos uma atitude positiva da personagem ("eles") para com o serviço.

Tendo em conta estes dois tipos de subjetividade, podemos verificar que, em certas ocasiões, poderá ser necessário identificar, juntamente com a subjetividade, o titular da opinião.

As técnicas de classificação de subjetividade podem ser baseadas em aprendizagem automática ou em léxicos. Um exemplo de classificação de subjetividade através de técnicas de aprendizagem automática é o de Pang and Lee (2004). Estes recolheram 5000 frases subjetivas relativas a avaliações de filmes e 5000 frases objetivas referentes a enredos de filmes para treinarem os modelos. A classificação de subjetividade foi executada em dois modelos automáticos diferentes (SVM e *Naive Bayes*), sendo que ambos foram alimentados com informação baseada em pares (por exemplo, que duas frases devem pertencer à mesma classe de subjetividade) e por item. Já Bravo-Marquez et al. (2014) classificaram subjetividade utilizando técnicas de aprendizagem baseadas em léxicos. Recolheram assim diversos atributos de vários léxicos, como por exemplo o somatório do valor das palavras positivas (SWP) e das palavras negativas (SWN) de um texto, que estejam presentes no vocabulário do SentiWordnet criado por Baccianella et al. (2010), para determinar se uma frase é subjetiva ou não.

2.3.4 *Classificação de Sentimentos*

A classificação de sentimentos é definida pelo processo de atribuição de uma conotação positiva ou negativa ao conteúdo subjetivo. É importante salientar que a classificação não é obrigatoriamente binária (positiva ou negativa), sendo que esta pode ser ternária (incluindo, por exemplo, a classe neutra) ou mesmo nominal. Existem vários modelos para a aplicação da classificação de sentimentos. A figura 4, adaptada de Medhat et al. (2014), apresenta graficamente esses modelos e as suas relações.

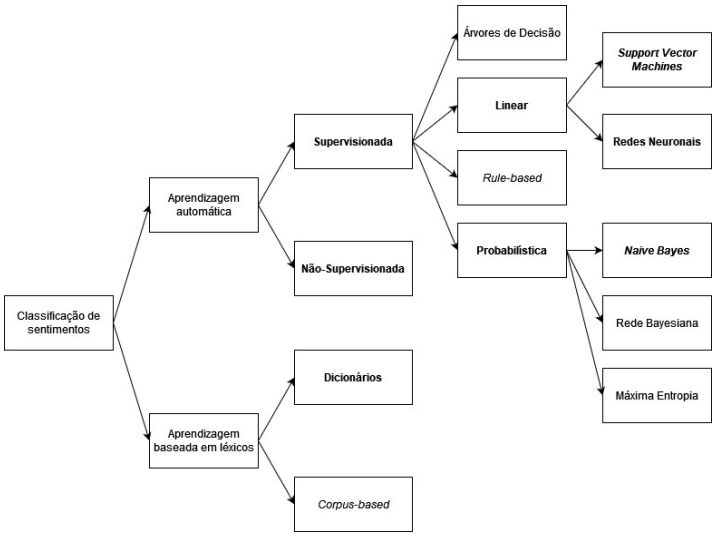

Figure 4: Modelos de classificação de sentimentos

Os modelos de classificação de sentimentos podem ser divididos em aprendizagem automática ou baseada em léxicos. A aprendizagem automática pode ainda ser dividida em aprendizagem supervisionada e não supervisionada. A aprendizagem baseada em léxicos pode ser dividida em aprendizagem baseada em dicionários

e baseada em corpus. Tradicionalmente, as técnicas utilizadas eram baseadas em léxicos, mas, com a introdução da inteligência artificial, a aprendizagem automática começou a ser amplamente explorada.

De momento, os métodos mais utilizados para classificar sentimentos são os métodos de aprendizagem automática, já que tendem a obter resultados melhores do que os métodos baseados em léxicos. Ainda assim, os métodos baseados em dicionários são algo frequentes, especialmente quanto utilizados para a criação de modelos híbridos. Os modelos híbridos são, como o próprio nome indica, modelos que utilizam ambas as abordagens automáticas e baseadas em léxicos.

Aprendizagem supervisionada

A aprendizagem supervisionada depende fortemente da existência de dados de treino classificados. Como já foi referido anteriormente, a aquisição de dados e o pré-processamento destes dados são fases importantíssimas para o bom funcionamento dos sistemas de análise de sentimentos, sendo que, para sistemas que implementem aprendizagem supervisionada, ainda mais importantes o são. Para treinar estes modelos, é importante recorrer a técnicas de extração e de seleção de atributos textuais presentes nos dados já processados. Algumas das técnicas mais comuns são:

- Presença de termos e a sua frequência – que possibilitam a atribuição de um valor, seja ele binário (o termo aparece ou não) ou nominal, para indicar a relevância de uma palavra ou conjunto de palavras num texto;
- *Parts of speech* (POS) – que permitem fazer a identificação de características textuais que indiquem a existência de sentimentos. Também podem ser consideradas técnicas de classificação de subjetividade. Por exemplo, a utilização de POS para identificação de adjetivos é bastante comum já que, tendencialmente, adjetivos exprimem emoções;
- Palavras de opinião ou frases – que fazem a deteção de palavras ou frases que são frequentemente utilizadas para expressar opiniões. Esta deteção pode ser extremamente simples, identificando palavras como “bom” ou “mau”, ou mais complexa, através da identificação de expressões mais singulares como, por exemplo: “Hoje estou com a telha.” significa que o titular da opinião está de mau humor (polaridade negativa);
- Negações – que nos permitem fazer a identificação de palavras negativas que, por norma, alteram a orientação da opinião. Por exemplo, na frase “Não gosto nada dele”, o termo “não gosto” é de polaridade negativa, ao contrário do termo “gosto” que é de polaridade positiva.

A presença e/ou frequência de termos em opiniões é a técnica utilizada com mais frequência para extração de características textuais. Isto deve-se, principalmente, à facilidade de a aplicar através de métodos automáticos. O modelo *“Bag-of-Words”* (BOW) e o modelo *“Term Frequency-Inverse Document Frequency”* (TF-IDF) são exemplos de técnicas que aplicam princípios de presença ou de frequência de termos em opiniões. No caso do modelo TF-IDF, cada palavra de uma frase de opinião é traduzida num valor estatístico que, no fundo, representa

a importância da palavra nessa mesma frase. O valor aumenta proporcionalmente ao número de palavras existentes na frase.

Para além das técnicas expostas anteriormente, novas abordagens têm vindo a surgir, como por exemplo abordagens baseadas em *word embedding*. Estas abordagens utilizam um método que consiste em representar palavras em vetores, sendo que palavras com um significado similar têm tendência a ser representadas por vetores mais próximos dentro do espaço vetorial que é composto por vetores representativos das palavras presentes no *corpus* utilizado. Mikolov et al. (2013) exploraram extensivamente este tema e concluíram a utilização de *word embedding* resulta numa melhoria significativa na performance de alguns modelos, associada a um custo computacional mais reduzido e, consequentemente, criaram o modelo *"Word2Vec"*. Em pesquisas posteriores, e como forma a facilitar tarefas de análise de sentimentos ao nível da frase e ao nível do documento, surgiram novas técnicas de *document embedding*. Le and Mikolov (2014) explicitaram uma metodologia onde parágrafos inteiros são representados por vetores, algo que é fortemente fundamentado tendo por base o método exposto por Mikolov et al. (2013), e que levou ao surgimento de múltiplos modelos, como por exemplo, o modelo *"Doc2Vec"*.

Após uma revisão da literatura sobre extração de características dos textos de opinião, e tendo em conta que estas características vão ser utilizadas como *input* para treinar os algoritmos de aprendizagem supervisionada, achamos que seria útil fazer, também, uma análise bibliográfica de cada um deles. Como podemos observar na figura 4, os métodos de aprendizagem supervisionada estão divididos em 4 classes principais: classificadores de árvores de decisão, lineares, probabilísticos e baseados em regras. Desta relação, o tipo de classificadores mais comumente utilizados para tarefas de classificação de sentimentos são os métodos lineares e probabilísticos. Dos métodos lineares e probabilísticos podemos identificar, no contexto da análise de sentimentos, como relevantes os seguintes algoritmos: *"Multinomial Naïve Bayes"*, *"Support Vector Machines"* e Redes Neuronais. Analisemos um pouco cada um deles.

Multinomial Naive Bayes

O algoritmo *"Multinomial Naive Bayes"* faz parte da família dos classificadores probabilísticos que se baseiam na aplicação do "teorema de Bayes" assumindo que existe uma total independência entre os atributos de cada exemplo num conjunto de dados. Esta assunção não representa de todo a realidade, daí a utilização do termo *"Naive"*, mas, usualmente, estes modelos demonstram uma eficácia bastante alta em tarefas de classificação, especialmente em tarefas de classificação aplicadas a conteúdo textual. Para além disso, esta assunção permite que a aprendizagem seja realizada de uma forma mais rápida e económica, já que cada atributo é tratado como um atributo independente, sem existir a necessidade de ter em consideração os parâmetros dos restantes atributos. Estas vantagens têm um peso relativamente maior em tarefas de classificação de opiniões, visto que a dimensão do espaço dos atributos utilizados nestas tarefas é regularmente extensa.

Segundo McCallum et al. (1998), o classificador *"Multinomial Naive Bayes"* permite captar informação como o número de ocorrências de uma palavra num documento, ignorando assim o seu contexto ou a sua posição no respetivo documento. Logicamente, este tipo de informação é muito semelhante à informação extraída de textos de opinião quando é utilizado o modelo "BOW" ou o modelo "TF-IDF", onde são criados vetores representativos da frequência/presença de termos num documento. Tendo em conta que o *"Multinomial Naive Bayes"* é perfeitamente

adaptável a este tipo de input, é natural que seja utilizado, à semelhança dos modelos "BOW" e "TF-IDF", com muita frequência.

Support Vector Machines

O algoritmo *"Support Vector Machines"* (SVM) é um modelo de classificação linear que tem como objetivo encontrar hiperplanos que permitam distinguir e classificar corretamente as classes do conjunto de dados alvo. Estes hiperplanos, que têm uma dimensão equivalente ao resultado de subtrair um ao número de atributos utilizadas para treinar o modelo, dividem o espaço de decisão de forma que seja possível separar um conjunto de dados tendo em conta a sua classe respetiva, sendo que, apesar de ser possível traçar vários hiperplanos que respeitem esta propriedade, o escolhido deve ser sempre representativo do hiperplano que garante a maior margem de distância entre os dados das diferentes classes. Para facilitar a compreensão do funcionamento do modelo, podemos supor o seguinte para um determinado problema:

- foram utilizados 3 atributos para treinar o modelo – x, y e z;
- o conjunto de dados tem 2 classes distintas – b (azul) e r (vermelho).

Tendo em conta o problema em questão, o hiperplano será um plano (duas dimensões) e não uma reta, já que o espaço de decisão é definido através do número de atributos (três) utilizadas no conjunto de dados. Na Figura 5 podemos observar um hiperplano definido para um espaço de decisão tridimensional:

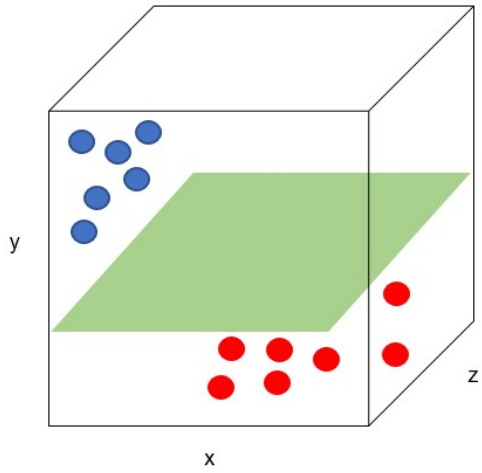

Figure 5: Exemplo de um hiperplano definido para um espaço de decisão tridimensional

Após uma breve introdução ao modelo, é necessário compreender o porquê deste modelo ser, teoricamente, ótimo para tarefas de classificação de sentimentos. Ora, segundo Joachims (1998), existem algumas evidências teóricas que justificam a boa performance dos modelos SVM quando aplicados a este tipo de tarefas, nomeadamente:

- a natureza dimensional dos atributos – usualmente quando lidamos com problemas de análise de sentimentos, o número de atributos utilizado para treinar os modelos costuma ser significativo, algo em concordância com o tamanho dos vocabulários em que os problemas estão inseridos. Quando trabalhamos com dados com uma grande quantidade de atributos, corremos o risco de introduzir *overfitting* num modelo, já que é simples os modelos encontrarem relações esporádicas entre os atributos. Sendo que o modelo SVM utiliza um mecanismo de regularização que previne *overfitting*, este acaba por suportar grandes espaços de decisão;

- a relevância dos atributos – uma forma de ignorar um grande espaço dimensional de atributos, é considerar que a maior parte delas é irrelevante. Este método retiraria a limitação apresentada anteriormente. Porém, em problemas que envolvem textos de opinião, retirar atributos pode significar uma perta significativa de informação, tendo já se comprovado que mesmo as atributos consideradas "menos relevantes" em problemas de classificação de conteúdo textual são fonte de um nível considerável de informação.

Redes Neuronais

Uma rede neuronal é um modelo de classificação linear inspirado no funcionamento dos sistemas neuronais biológicos, como é o caso do cérebro humano. No geral, as redes neuronais são compostas por múltiplas camadas de neurónios que são constituídas por uma camada de entrada (*input layer*), uma camada de saída (*output layer*) e uma ou mais camadas intermédias (*hidden layers*). Cada nodo de uma camada está conectado aos nodos da camada seguinte, sendo que esta ligação tem um peso e um *threshold* associado, o qual permite controlar a ativação ou não de um nodo, ou seja, a passagem de dados entre camadas.

Existem vários tipos de redes neuronais como é o caso das redes neuronais convolucionais e das redes neuronais recorrentes, sendo que as redes neuronais recorrentes têm vindo a ser exploradas intensivamente em problemas de classificação de conteúdo textual devido à sua grande adaptabilidade a conjuntos de dados sequencias, isto é, conjuntos de dados que tenham uma ordenação definida. No contexto da análise de sentimentos, é oportuno realçar os modelos *"sequence-to-vector"*, que recebem um *input* sequencial, como uma opinião, e retornam um vetor, que neste caso pode ser representativo do quão boa ou má essa opinião é. As redes neuronais recorrentes têm, ainda assim, algumas limitações, como por exemplo:

- o demorado processo de aprendizagem;

- pouca adaptabilidade a sequências longas, o que é uma característica comum em opiniões.

Estas limitações ocorrem, não só, mas principalmente, em virtude da necessidade deste tipo de modelos receberem o seu input sequencialmente, retirando assim a possibilidade de paralelizar o treino dos modelos, inutilizando assim o poder de processamento das atuais unidades de processamento gráfico (GPU). Surgem assim as redes neuronais *"transformer"* (TNN) Vaswani et al. (2017). A principal diferença entre estas e as redes neuronais recorrentes é o facto das TNN receberem os dados de uma forma paralelizada.

As redes neuronais *"transformer"* são compostas por dois componentes principais: um *encoder* e um *decoder*. Para simplificar a explicação do funcionamento dos componentes, vamos considerar um problema de tradução entre português e inglês.

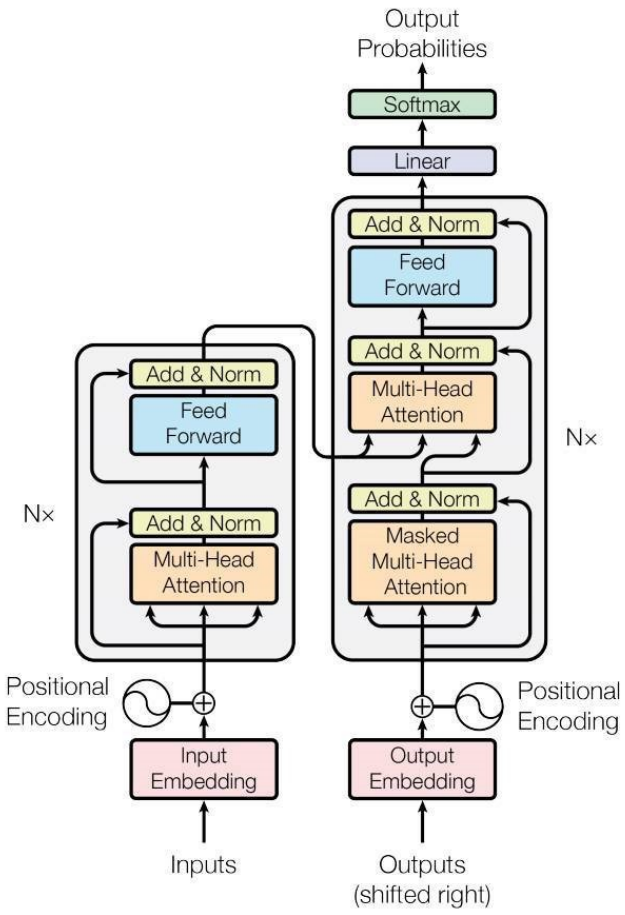

Figure 6: Arquitetura das redes neuronais *"transformer"* – figura extraída de Vaswani et al. (2017))

1. Inicialmente, o *encoder* recebe, em simultâneo, várias palavras em português que são convertidas para vetores tendo em conta o espaço onde estão inseridas, espaço este que é definido pelo conjunto de palavras que têm um significado semelhante (*"input embedding"*).

2. De seguida são criados vetores de contexto que são a soma dos vetores provenientes da camada de *"input embedding"* com os vetores de posição das palavras na frase (*"positional encoding"*). Esta fase permite distinguir palavras iguais que têm significados diferentes quando inseridas numa determinada frase;

    a) De seguida, estes vetores são passados ao *"encoding block"*, que contém uma camada *"multi-head attention"* e uma camada *"feed forward"*. A camada *"multi-head attention"* tem como objetivo computar vetores de "atenção" para cada palavra, sendo que a atenção é medida pelas relações que uma palavra tem com as restantes palavras da mesma frase. Regularmente, estes vetores de "atenção" são extremamente focados na própria palavra, contendo assim pouca ou nenhuma

informação pertinente. Por essa mesma razão, são computados vários vetores de "atenção" para a mesma palavra e, de seguida, todos os vetores são concatenados para que seja possível recolher uma quantidade significativa de informação.

b) Já a camada de *"feed forward"* tem um objetivo mais simples, que consiste simplesmente em transformar os vetores provenientes da camada *"multi-head attention"* para vetores que sejam facilmente compreendidos pelas camadas seguintes.

3. Semelhante ao processo de *encoding* apresentado nos pontos anteriores, é o processo realizado no decoder, que recebe as palavras portuguesas introduzidas no encoder, já traduzidas para inglês. O processo do decoder é muito semelhante ao do *encoder*, passando por uma camada de *"embedding"*, *"positional encoding"* e *"multi-head attention"*, e retornando um vetor de "atenção" para cada palavra;

4. Os vetores provenientes do *encoder*, juntamente com os vetores provenientes do *decoder*, são utilizados como *input* para o resto do processo de *decoding*, que permitirá inferir o quão semelhantes são as palavras portuguesas das palavras traduzidas para inglês.

No que diz respeito ao *decoder*, existem muitos outros detalhes que aqui não foram mencionados. Ainda assim, para o contexto da análise de sentimentos, é importante manter o foco no *encoder* que, como pode ser depreendido, é o bloco que realmente tem como objetivo decifrar o significado das palavras e compreender a linguagem utilizada. Se executarmos este processo de *encoding* várias vezes (para uma aprendizagem mais aprofundada) obtemos o BERT – *Bidirection Encoder Representation from Transformers* Devlin et al. (2018).

O BERT tem vindo a ser utilizado com recorrência para executar tarefas de análise de sentimentos, e está dividido, essencialmente, em duas tarefas principais:

- Pré-treino, que consiste na fase de aprendizagem de linguagem por parte do algoritmo;
- *Fine tuning*, que tem como objetivo adaptar o algoritmo ao problema que queremos resolver, como obter uma classificação negativa ou positiva para uma opinião, por exemplo.

A utilização comum do BERT deve-se, principalmente, aos bons resultados que tem alcançado e a sua adaptabilidade às diferentes tarefas de processamento de linguagem textual.

Aprendizagem não supervisionada

Frequentemente, na análise de sentimentos é difícil encontrar dados que subscrevam os requisitos necessários para implementar métodos de aprendizagem supervisionada. Efetivamente, recolher opiniões não é algo complexo, mas recolher opiniões corretamente classificadas e que estejam classificadas tendo em conta o objetivo final do sistema de análise de sentimentos é árduo. De forma a contrariar estes obstáculos, podem ser utilizados métodos não supervisionados, que permitem classificar opiniões sem recorrer a dados para treinar modelos.

As técnicas mais comuns de aprendizagem não supervisionada são métodos baseados em léxicos. O processo usual de aprendizagem baseada em léxicos consiste em recolher um léxico pré-definido para determinar a

polaridade geral de um documento, sendo que um léxico é usualmente composto por palavras e a sua respetiva polaridade. Assim, podemos dividir as técnicas de aprendizagem não supervisionada em duas categorias principais: aprendizagem baseada em dicionários e aprendizagem baseada em *corpus*. A aprendizagem baseada em dicionários é a mais utilizada devido à sua inerente facilidade em generalizar os problemas de classificação de textos, e baseia-se na aprendizagem da orientação semântica de uma palavra, tendo em conta a orientação semântica das palavras, presentes num respetivo dicionário, semelhantes a si, tanto ao nível semântico como ao nível linguístico.

2.4 DESAFIOS E OPORTUNIDADES

O tema da análise de sentimentos é ainda um tema em exploração, pelo que, por vezes, são identificados novos desafios e problemas associados a implementações já existentes. É importante compreender quais são esses desafios, para que se possa depreender quais são as limitações das implementações existentes. A exploração de soluções para estes desafios, é também um enorme contributo para o tema da análise de sentimentos, abrindo assim espaço para novas descobertas e oportunidades. De seguida, apresentamos os principais desafios referentes à área da análise de sentimentos, bem como algumas implementações de soluções para alguns deles. De referir, então:

- Contexto – frequentemente, os sistemas de análise de sentimentos são construídos tendo por base um contexto específico, como por exemplo avaliações de filmes ou telemóveis. Estes sistemas, quando aplicados a contextos diferentes do original, revelam uma enorme perda de eficácia e qualidade. Para além disso, algumas palavras de sentimento têm uma conotação diferente quando aplicadas em contextos diferentes. Por exemplo, tomemos em consideração as frases "As sirenes desta ambulância são silenciosas" e "Durante a noite, o meu prédio é silencioso". O facto de as sirenes serem silenciosas é algo negativo, já que estas devem ser ouvidas com clareza. Já o prédio ser silencioso durante a noite é algo vantajoso.

- Negações – a não identificação de negações é um problema comum em sistemas de análise de sentimentos. Este problema tem origem no efeito que as negações têm na alteração da polaridade real de uma opinião. Atualmente, alguns sistemas de análise de sentimentos já utilizam técnicas para contrariar este problema, como a identificação de palavras de negação.

- Frases objetivas – usualmente as opiniões são expressas através de subjetividade. Ainda assim, existem algumas situações, raras, nas quais o conteúdo objetivo expressa sentimentos. Estas situações normalmente identificam-se através da exposição de factos indesejáveis relativamente a uma entidade. Por exemplo, veja-se a frase "Tenho este telemóvel há pouco tempo e os botões já estão estragados".

- Frases condicionais – como já foi referido anteriormente, os sistemas de análise de sentimentos baseiam-se com frequência em palavras de sentimento para indicar presença de subjetividade. Isto pode gerar um problema, especialmente se essas palavras estiverem incluídas em frases condicionais, nas quais não

expressam qualquer tipo de emoção. Inevitavelmente, a análise destas frases representa uma tarefa mais difícil e, por isso, já surgiram estudos relacionados com a forma de as abordar no contexto da análise de sentimentos. Narayanan et al. (2009), por exemplo, estabeleceram diferentes combinações de atributos extraídos dos dados de treino classificados, como as palavras de negação e os conectores condicionais. De seguida, aplicaram essas combinações a diferentes tipos de estratégias de classificação, como por exemplo, baseadas na totalidade da frase ou apenas na cláusula condicional - "Se este telemóvel for bom, compro-o" é um exemplo desses casos.

- Sarcasmo - a deteção de sarcasmo em formato textual é um grande desafio para o ser humano, e este desafio torna-se ainda maior quando a tarefa é realizada por sistemas computacionais. Por exemplo, veja-se o significado de uma frase como esta: "Que jogo bem feito, nem consigo ver a equipa adversária".

2.5 SISTEMAS E UTILIDADE

2.5.1 *Utilidade*

Atualmente, a análise de sentimentos pode influenciar positivamente um vasto número de áreas aplicacionais. Na realidade, esta capacidade abrangente não passa despercebida, até porque empresas como a Microsoft e a Google já desenvolveram as suas próprias soluções de análises de sentimentos. Sendo assim, algumas das principais aplicações da análise de sentimentos, são Ravi and Ravi (2015):

- Previsão do mercado de ações – apesar de ser um mercado extremamente especulativo, é possível fazer algumas previsões, já que este pode ser influenciado por imensos e diversos fatores. Um deles pode ser as opiniões dos consumidores, isto porque se os consumidores de um ou mais produtos de uma empresa não estiverem satisfeitos, a probabilidade do número de vendas desses produtos diminuir é maior, logo, impactará negativamente o valor das ações da empresa em questão. Bollen et al. (2011) testaram o impacto do estado de espírito do Twitter nos dados financeiros dados pelo índice de mercado *Dow Jones Industrial Average* (DIJA). A análise foi feita através de duas ferramentas de deteção de humor – *Opinion Profile* (OF) e *Google-Profile of Mood States* (GPOMS) – sendo que, inicialmente, foi testada a eficácia das ferramentas através da comparação dos valores dos índices destas com eventos de importância, como as eleições presidenciais e o dia de ação de graças. Tendo em conta os resultados positivos, foi depois observada a relação preditiva entre o DIJA e as ferramentas OF e GPOMS, e comprovou-se que, de facto, existe uma relação forte entre algumas dimensões do GPOMS, como a calma, com o DIJA.

- Previsão do sucesso da bilheteira – à semelhança do mercado de ações, as opiniões dos consumidores representam um peso no sucesso da bilheteira de, por exemplo, um filme. Neste caso, o peso das opiniões é muito superior ao peso que elas têm no mercado de ações, isto porque os filmes dependem fortemente daquilo que o público geral acha deles. Rui et al. (2013) estudaram o peso das opiniões, retiradas de *tweets*, no valor das vendas de filmes. Inicialmente utilizaram métodos de aprendizagem supervisionada para determinar a subjetividade e a classificação dos *tweets*. Este estudo mostrou que, por exemplo, um

aumento de 1% no número de *tweets* positivos sobre um filme, aumenta a sua receita em cerca de 125000 dólares na semana consequente.

- Marketing inteligente – ajuda a determinar, por exemplo, o sucesso de uma campanha de anúncios tendo em conta as opiniões dos consumidores. Também serve para criar anúncios personalizados tendo em conta as opiniões de um utilizador sobre os produtos de uma empresa. É realmente um recurso muito importante atualmente, já que o marketing é uma das ferramentas mais poderosas para atrair novos clientes e para garantir a permanência dos consumidores já clientes. Li and Li (2013) criaram um sistema de marketing inteligente através da observação de opiniões em micro-blogues. Este sistema, como mencionado no artigo, é um sistema de monitorização que permite compreender tendências de mercado, através da observação das flutuações dos sentimentos sobre um tópico em particular. Desta forma, permite tomar as decisões mais corretas.

- Inteligência Competitiva – algo semelhante ao marketing inteligente, permite extrair e visualizar comparações entre produtos concorrentes a partir de opiniões dos consumidores. Também apoia a análise de potenciais riscos e o design de novos produtos. Resumidamente, dá a possibilidade às empresas de estar um passo à frente da sua concorrência e também de melhorar efetivamente o(s) produto(s) ou serviço(s) comercializados por estas. Kang and Park (2014) analisaram, tendo em conta várias dimensões, opiniões de consumidores acerca de aplicações móveis como o Facebook, o Skype e o Twitter. Alguns resultados deste estudo, mostram uma necessidade de investimento nos serviços de pesquisa e toque, para aumentar o nível de satisfação dos consumidores. Outros resultados mostram que o Twitter é a aplicação com o nível mais alto de satisfação dos utilizadores, sendo que deve ser considerada uma referência em praticamente todas as dimensões. Como podemos verificar, informação como esta é extremamente importante para empresas que querem competir com qualidade no mercado.

- Sistemas de recomendação – os sistemas de recomendação são, atualmente, extremamente explorados. Todos os dias vemo-los em funcionamento, em plataformas como a Netflix ou o Youtube. São considerados ferramentas poderosas, já que criam uma experiência dos sistemas e serviços mais agradável para os utilizadores. Estes sistemas podem ser construídos através das opiniões dos utilizadores, recolhendo os seus sentimentos relativamente a diferentes tópicos para depreender os seus gostos. Li and Shiu (2012) desenvolveram um sistema de recomendações que disponibiliza uma lista de utilizadores a que um anúncio deve ser providenciado, tendo em conta as preferências demonstradas por estes em redes sociais.

Para além das áreas mencionadas anteriormente, existem muitas outras áreas onde a análise de sentimentos pode ter um enorme efeito, como:

- o bem-estar da população, já que pode representar um meio para avaliar o bem-estar de um indivíduo ou conjunto de indivíduos tendo em conta um determinado assunto;

- a previsão de eleições políticas, na qual, analisando a opinião da população relativamente a, por exemplo, um candidato, é possível depreender a sua popularidade e, consequentemente, o seu resultado nas eleições em causa.

2.5.2 *Sistemas Reais*

Como podemos depreender a partir da secção anterior, a análise de sentimentos tem grande aplicação em váris domínios que, para além de envolver um vasto e diverso número de domínios, contém tópicos de alta relevância do ponto de vista económico e social. Sendo assim, o processo de análise de sentimentos é comumente integrado em sistemas que explorem estes tópicos, como por exemplo:

- O Moodlens Zhao et al. (2012) que é um sistema que analisa os sentimentos presentes em *tweets* chineses. Esta análise é baseada em *emoticons* e consegue detetar quatro tipos de sentimentos: zangado, enojado, feliz e triste. Para além disso, o sistema inclui a capacidade de processar e classificar *tweets* em tempo real, permitindo assim a monitorização da flutuação dos sentimentos.

- O VISA Duan et al. (2012) que é um sistema visual de análise de sentimentos. O seu objetivo é demonstrar o conteúdo do *corpus* textual de diferentes tópicos e tendências. Basicamente consiste num sistema que não só incluí o processo da análise de sentimentos, mas também uma interface constituída por um *dashboard* que apresenta as flutuações dos sentimentos sobre diferentes temas e tendências em tempo real.

- O MSAS Chamlertwat et al. (2012) (*Micro-blog Sentiment Analysis System*) que é um sistema de análise de sentimentos capaz de analisar automaticamente opiniões de consumidores provenientes do Twitter. Este sistema utilizou, como caso de estudo, 100000 publicações relacionadas com *smartphones*, e conseguiu retirar informação importante no que diz respeito aos sentimentos dos consumidores relativamente a características do produto, especificamente relacionadas com o ecrã, as aplicações e a câmara.

- O Crystal Kim and Hovy (2007) que é um sistema de análise de sentimentos com capacidade para prever eleições através das opiniões postadas por utilizadores num *website* de previsão de eleições. Utiliza métodos supervisionados para prever as eleições, recorrendo a dados retirados das opiniões por métodos de generalização de atributos. O Crystal obteve uma eficácia de 81.68% a prever eleições futuras.

3

PREPARAÇÃO DE DADOS

3.1 O PROCESSO DE PREPARAÇÃO DE DADOS

Ainda que a aquisição de grandes quantidades de dados seja, atualmente, simples, nem sempre estes conjuntos de dados estão estruturados de uma forma adequada aos nossos processos. Usualmente, verificam-se vários tipos de inconsistências nos dados, como, por exemplo, a ocorrências de valores nulos ou a desformatação de dado. Assim, é necessário fazer uma preparação e uma análise cuidada dos dados que permita criar condições para que estes dados possam ser utilizados no sistema alvo. A este processo é normalmente atribuída a denominação de pré-processamento. Na Figura 7 podemos observar uma representação típica da fase de pré-processamento de dados:

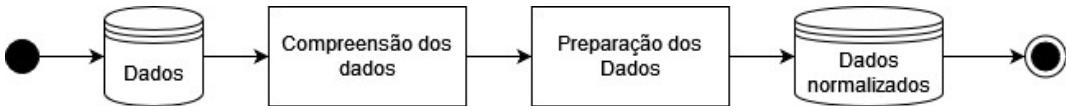

Figure 7: Pré-processamento típico

Na realidade, o esquema é uma versão baseada na metodologia padrão aplicada em processos de mineração de dados, uma metodologia independente do setor industrial onde é aplicada e das tecnologias utilizadas – *Cross Industry Standard Process for Data Mining* (CRISP-DM) Wirth and Hipp (2000). Tendo em conta a sua vasta adaptabilidade, a metodologia CRISP-DM será a base para as soluções apresentadas nas secções seguintes.

A fase de preparação de dados é de imensa importância para o bom funcionamento dos sistemas de análise de sentimentos e, por isso, devemos tê-la em grande consideração, planeando meticulosamente cada uma das suas subfases integrantes. Importante mencionar que, para além de garantir um melhor funcionamento das soluções de análise de sentimentos, um pré-processamento bem estruturado resulta num melhor aproveitamento do tempo e de recursos, já que afasta a necessidade de voltar atrás depois de já estarem elaborados, construídos e executados os modelos de análise que, por si só, dependem fortemente da fase de preparação de dados.

3.2 DEFINIÇÃO E ESTRUTURAÇÃO DOS ELEMENTOS DOS DADOS

O primeiro passo do processo de preparação de dados é fazer a definição correta dos elementos que devem estar presentes no conjunto final de dados recolhidos. Para que isto seja realizado corretamente, é necessário compreender totalmente qual é o objetivo final do sistema ou da solução que estamos a desenvolver.

Olhando especificamente para o caso do trabalho desta dissertação, o objetivo principal da solução baseada em análise de sentimentos é retirar um valor numérico compreendido entre 1 e 5 que classifique a opinião de um utilizador, tendo em conta o sentimento que nela está expresso. Ora, para um melhor entendimento do que estes valores representam em termos de sentimentos expressos numa opinião, podemos considerar o seguinte:

- o valor 1 representa uma opinião com uma polaridade muito negativa;
- o valor 2 representa uma opinião com uma polaridade negativa;
- o valor 3 representa uma opinião com uma polaridade neutra;
- o valor 4 representa uma opinião com uma polaridade positiva;
- o valor 5 representa uma opinião com uma polaridade muito positiva.

Assim, o primeiro passo é garantir que os dados contêm uma coluna com um valor compreendido entre 1 e 5, e que este represente o sentimento expresso na respetiva opinião.

Em segundo lugar, é necessário compreender que tipo de opiniões são fulcrais para o sistema. Frequentemente, e como já foi referido anteriormente, as soluções baseadas em análise de sentimentos deparam-se com um problema: o contexto das opiniões. Na realidade, os modelos têm dificuldade em classificar opiniões quando o contexto delas varia com frequência. Isto pode acontecer, por exemplo, quando algumas palavras têm uma conotação diferente dependendo do contexto. Tendo isto em conta, e sabendo que o sistema para o qual esta solução está a ser desenvolvida é baseado em inquéritos sobre o funcionamento de uma plataforma de *e-learning*, seria útil recolher dados que estejam inseridos no contexto educativo para melhorar a eficácia dos modelos.

Finalmente, já foi também aqui referido o problema referente ao idioma no qual as opiniões foram redigidas. Como sabemos, a estrutura sintática do texto e as palavras de opinião variam significativamente de idioma para idioma. Logo, o que pode ser considerado um comentário negativo num idioma, pode ter outro tipo de sentimento associado quando traduzido para outro idioma. Em contrapartida, as ferramentas existentes para processamento textual (tarefas de NLP) estão significativamente otimizadas para textos escritos em inglês, dificultando assim a fase de pré-processamento quando as opiniões recolhidas estão num idioma diferente do inglês. Sendo que já existe uma limitação considerável aos conjuntos de dados que podem ser recolhidos, imposta pelos requisitos definidos anteriormente, neste trabalho serão recolhidas opiniões redigidas em inglês, uma vez que as opiniões expressas neste idioma são mais comuns e em maior número. Além disso, facilitam substancialmente a fase de pré-processamento.

Após uma procura exaustiva de dados que contivessem os requisitos referidos anteriormente, foi possível concluir que, apesar de já existirem vários conjuntos de dados contendo opiniões anotadas com um valor entre 1

e 5, não existe nenhum conjunto de dados pré-definido que contenha opiniões de utilizadores sobre plataformas educacionais anotadas com um valor entre 1 e 5. Ainda assim, durante o processo de procura, foi possível encontrar uma plataforma de cursos online que contém uma vasta quantidade de opiniões sobre os seus cursos e certificados. Na realidade, esta plataforma contém opiniões inseridas no contexto educacional, escritas em inglês, e anotadas com uma classificação entre 1 e 5. Como a plataforma em causa não fornece qualquer tipo de API que auxilie o processo de recolha de dados, foi desenvolvido um *web scrapper* para extrair automaticamente as opiniões e as suas respetivas classificações. Na figura 8 podemos observar o processo geral de recolha de dados que foi implementado:

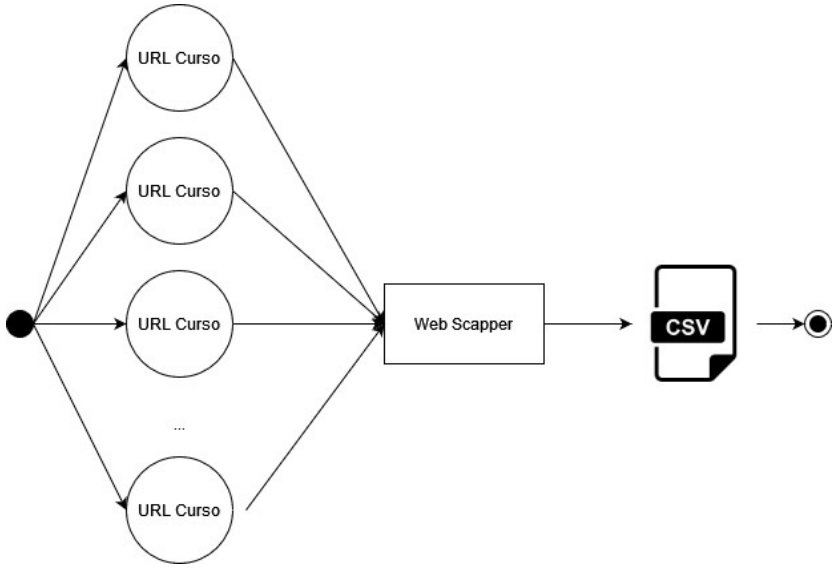

Figure 8: O processo de recolha de dados implementado

No processo de recolha de dados realizado foram angariadas opiniões sobre vários cursos da plataforma online referida, o que permitiu construir um conjunto de dados de opiniões sobre o funcionamento de cursos sensivelmente comprido e robusto. Durante o processo de *web scrapping*, foi realizada uma verificação do idioma da opinião de forma a recolher comentários apenas em inglês.

Para além de todo este processo, foi adicionada uma nova coluna ao conjunto de dados que contém a descrição da polaridade da opinião, isto é, as opiniões classificadas com 1 ou 2 foram descritas como negativas, as opiniões classificadas com 3 foram descritas como neutras, e as opiniões classificadas com 4 ou com 5 foram descritas como positivas. Esta coluna foi adicionada para possibilitar a comparação entre os modelos desenvolvidos neste trabalho de dissertação com outros modelos de análise de sentimentos mais célebres, já que estes, por norma, classificam opiniões como positivas e negativas, podendo também incluir a classe “neutra”.

3.3 ANÁLISE DOS DADOS DISPONÍVEIS

O passo seguinte no processo de preparação de dados é a análise do conjunto de dados recolhido. Este processo permite observar características relevantes do conjunto de dados, que ajudarão a compreender quais os processos de transformação que terão que ser aplicados nas fases posteriores. Por exemplo, observando a tabela 3, podemos verificar que o conjunto de dados contém 277287 opiniões de utilizadores. Para além disso, cada opinião está associada à sua polaridade e classificação respetiva. O conjunto de dados não contém valores nulos, o que significa que o *web scrapper* desenvolvido funcionou corretamente.

Texto da Opinião	Estrelas	Polaridade
Not a comprehensive course.	1	*negative*
Please cancel this course. I want to opt out of it	1	*negative*
This course is good for just theoretical understanding of the subject. But for practical implementation it is too hard to do.	3	*neutral*
The course is a good balance between learning key concepts and doing coding, the coding being optional. The phrasing of quiz questions and answers were sometimes confusing.	4	*positive*
Very good introductory course, I highly recommend it to anyone looking to get a flavour of the methods behind the recent advances in AI without going into super-technical details.	5	*positive*

Table 3: Extrato de algumas linhas do conjunto de dados

Número de linhas	Número de colunas
277287	3

Table 4: Dimensões do conjunto de dados

De seguida é necessário analisar o conjunto de dados em termos de balanceamento, isto é, se existe uma ou mais classes com uma predominância muito superior à das outras classes no conjunto de dados recolhido. Este tipo de situações é extremamente comum em soluções de análise de sentimentos, isto porque, tendencialmente, os utilizadores publicam mais opiniões com uma conotação positiva do que propriamente negativa. Para além disso, quando isto acontece em soluções de aprendizagem automática, os modelos são usualmente fortemente influenciados, criando uma certa tendência de favorecimento para as classes mais predominantes. Por essa mesma razão, é necessário observar métricas, como a precisão, que nos permitam tirar conclusões sobre a influência das classes predominantes nos modelos. Tendo em conta que a análise de sentimentos é um processo que frequentemente envolve aprendizagem automática, retirar este tipo de informação na fase de análise dos dados é fulcral.

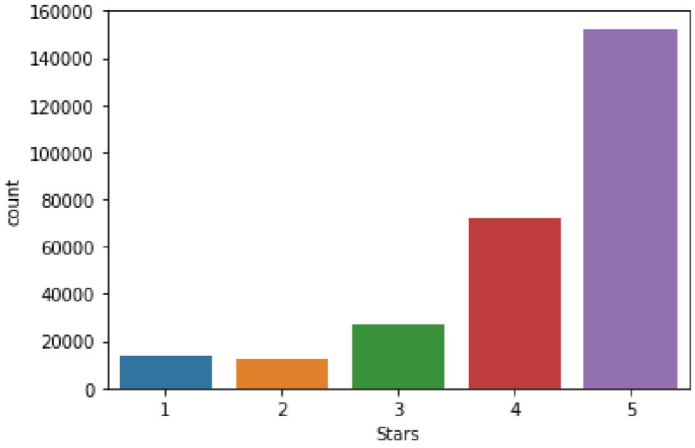

Figure 9: Número de opiniões por classificação

Como podemos verificar, o conjunto de dados com que trabalhámos não é exceção a outros conjuntos de dados usados em tarefas de análise de sentimentos, já que existe uma enorme predominância das duas classes com polaridade positiva (4 e 5) comparativamente às classes com polaridade neutra e negativa. Sendo assim, poderá ser necessário balancear o conjunto de dados na fase de treino e teste dos modelos.

Finalmente, sabendo que a eficácia dos modelos de análise de sentimentos depende muito do vocabulário derivado das opiniões recolhidas, isto porque, usualmente, estes vocabulários contêm uma grande quantidade de palavras não relevantes à análise de sentimentos, devemos analisá-lo atenciosamente.

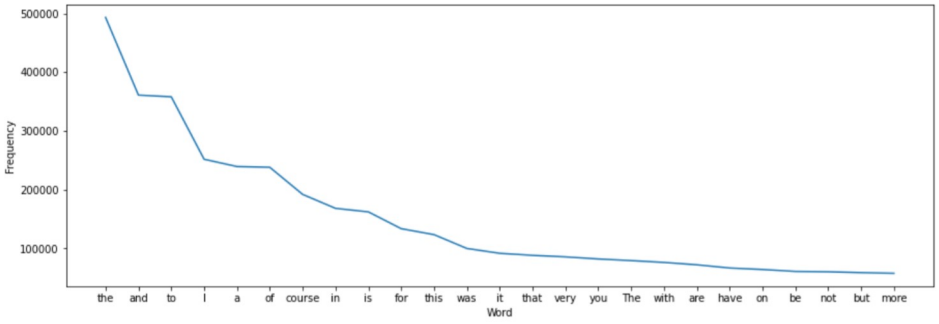

Figure 10: Número de ocorrências das 25 palavras mais comuns

Número de palavras
190493

Table 5: Número de palavras que ocorrem no conjunto de dados

Tendo em conta informação da tabela 5, é possível inferir que o conjunto de dados contém 190493 palavras diferentes. Para além disso, se olharmos para a figura 10, é possível verificar que as palavras "*the*", "*and*" e "*to*"

são as 3 palavras mais frequentes. Na realidade, esta informação é mais relevante do que aquilo que parece, já que estas 3 palavras representam termos que não são valorizados num sistema de análise de sentimentos, isto porque não contêm qualquer tipo de valor sentimental associado a eles. Tal circunstância ajuda-nos a tomar uma decisão na fase de tratamento e transformação dos dados, já que, utilizando processos como a remoção de "*stop words*", podemos reduzir imenso a complexidade dos dados, facilitando assim a execução dos modelos de aprendizagem automática.

3.4 TRATAMENTO E TRANSFORMAÇÃO DOS DADOS

Depois de se analisar extensivamente o conjunto de dados recolhido, é necessário fazer algumas transformações. Ora, tendo em conta que os dados recolhidos são dados com conteúdo textual, esta tarefa é praticamente indispensável, sendo que em alguns textos verifica-se, por exemplo, uma destruturação sintática. Na figura 11, podemos observar o processo desenhado para a fase de tratamento e transformação de dados. Nas secções seguintes explicaremos o porquê da escolha deste processo, o porquê da escolha de cada uma das etapas do processo (através da análise de alguns exemplos) e o seu efeito no conjunto de dados final.

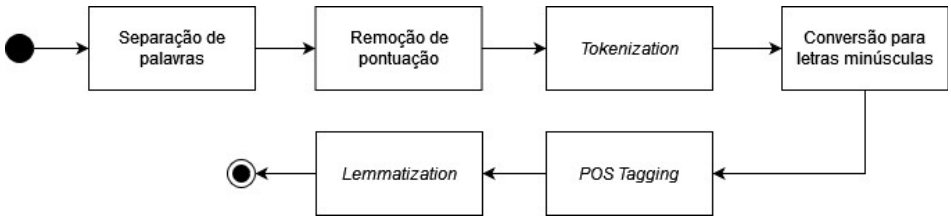

Figure 11: Processo de transformação dos dados

De notar que este processo incluí algumas das etapas mencionadas no capítulo 2, como por exemplo o processo de *Tokenization*. Ainda assim, muitas das etapas aqui utilizadas não foram mencionadas no capítulo 2, já que não são as mais comumente utilizadas. Muitas delas foram escolhidas especificamente para que os dados estejam processados de forma a que o seu conteúdo cumpre os requisitos necessários para a a execução do processo de classificação de sentimentos que será apresentado no capítulo 4.

3.4.1 *Separação de palavras e remoção da pontuação*

Um dos principais problemas no que diz respeito à qualidade dos dados referentes a opiniões, é a sua estrutura sintática. Sabendo que as opiniões são registadas em plataformas nas quais não é necessário qualquer tipo de estrutura de escrita, o formato da escrita é usualmente descuidado. Observemos os seguintes exemplos:

1. *Very good introductory course ,very well designed and professors explaination is very easy to understand .Go for it guys !Happy learning !!!!Sonic Somanna PK*

2. *I simply loved the course. I've been working with MachineLearning, but I didn't understand much about DeepLearning - this course helped me a lot to get started in this new research area.*

Na realidade, ainda que estas frases de comentários estarem incorretamente redigidas ao nível sintático, o ser humano consegue facilmente decifrar qual era a real intenção do utilizador. Infelizmente, isto não acontece com computadores, já que estes entendem os casos sublinhados como uma única palavra. Sendo assim, foi desenvolvido um pequeno programa que permite eliminar estas ocorrências, colocando espaços entre elas, separando assim corretamente as palavras. Na tabela 6 podemos ver um exemplo da aplicação da separação das palavras de um texto de uma opinião. Durante toda a explicação do processo de tratamento e transformação de dados, a mesma opinião será utilizada para facilitar a compreensão do *pipeline* criado.

Opinião	Separação de palavras
A very fine tuned Course,used as a warm up course for deep learning,highly recommended	*A very fine tuned Course , used as a warm up course for deep learning , highly recommended*

Table 6: Exemplo de aplicação de separação de palavras

Através da figura 12, podemos ver que o número de palavras no vocabulário reduziu imenso após a separação das palavras, o que significa que muitos destes casos ocorriam no conjunto de dados inicial.

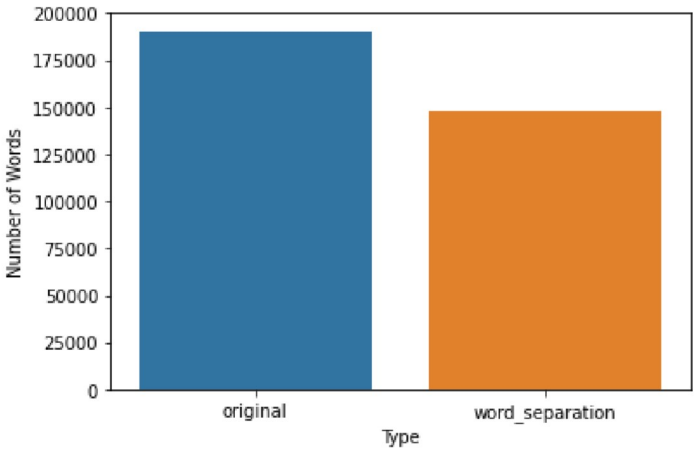

Figure 12: Tamanho do vocabulário após a separação de palavras

Após a separação de palavras, que, por si só, ajudou a reduzir a complexidade do sistema, foi também útil remover a pontuação das opiniões para facilitar a compreensão dos modelos de análise de sentimentos aplicados neste trabalho de dissertação. Ainda que se perca a estrutura sintática das opiniões, a pontuação é usualmente inútil no que diz respeito à análise de sentimentos, especialmente tendo em conta o contexto no qual estas opiniões estão inseridas e onde, raramente, são utilizados *emoticons* para expressar sentimento, ao contrário do que acontece, por exemplo, no Twitter.

Opinião	Remoção de pontuação
A very fine tuned Course , used as a warm up course for deep learning , highly recommended	*A very fine tuned Course used as a warm up course for deep learning highly recommended*

Table 7: Exemplo de aplicação de remoção de pontuação

3.4.2 *Tokenization e conversão para minúsculas*

A transformação de um comentário em *tokens* permite analisar individualmente as palavras de um comentário. O método mais utilizado para *tokenization* consiste na separação do texto por espaços. Tendo em conta que foi feita uma separação prévia das palavras e uma remoção da pontuação dos comentários, este método é eficiente para transformar a frase referente à opinião em *tokens*. Sendo assim, a função "word_tokenize" disponibilizada pela biblioteca NLTK foi utilizada para realizar o processo de tokenization. A tabela 8, apresenta o resultado do processo de tokenization aplicado a uma opinião.

Opinião	*Tokenizer*
A very fine tuned Course used as a warm up course for deep learning highly recommended	*["A", "very", "fine", "tuned", "Course", "used", "as", "a", "warm", "up", "course", "for", "deep", "learning", "highly", "recommended"]*

Table 8: Exemplo de aplicação de um *tokenizer*

Após o processo de *tokenization*, foi efetuada uma transformação de todas as palavras do vocabulário para palavras minúsculas. Na tabela 9 podemos ver um exemplo de conversão de todas as palavras de uma opinião para minúsculas:

Opinião	Transformação para minúsculas
["A", "very", "fine", "tuned", "Course", "used", "as", "a", "warm", "up", "course", "for", "deep", "learning", "highly", "recommended"]	*["a", "very", "fine", "tuned", "course", "used", "as", "a", "warm", "up", "course", "for", "deep", "learning", "highly", "recommended"]*

Table 9: Exemplo de aplicação de uma transformação para minúsculas

Mais uma vez, o objetivo deste processo foi reduzir a complexidade do sistema. Neste caso, o processo de conversão para letras minúsculas permite uniformizar ainda mais, não só o vocabulário como um todo, mas também cada comentário, o que se revela muito importante, sobretudo em abordagens que utilizam modelos supervisionados. Para compreendermos melhor a importância deste passo, vamos utilizar um modelo supervisionado baseado em BOW como referência. Ora, o modelo BOW produz um vetor do tamanho do

vocabulário, em que cada elemento do vetor representa uma palavra presente no vocabulário e, para cada opinião, verifica se a palavra está presente na opinião ou não. Considerando a opinião utilizada anteriormente, estes seriam os vetores gerados utilizando *bag of words* antes da conversão para minúsculas e depois da conversão para minúsculas (Tabela 10 e 11).

A	very	fine	tuned	Course	used	as	a	warm	up	course	for	deep	learning	highly	recommended
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 10: Vetor BOW antes da transformação para minúsculas

a	very	fine	tuned	course	used	as	warm	up	for	deep	learning	highly	recommended
1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 11: Vetor BOW após transformação para minúsculas

Como podemos verificar, ocorreu uma diminuição do vocabulário que, consequentemente, diminui significativamente o tamanho dos vetores, facilitando a capacidade de aprendizagem e convergência dos modelos supervisionados. Na figura 13, podemos observar claramente o efeito que o processo de transformação para minúsculas tem no tamanho do vocabulário.

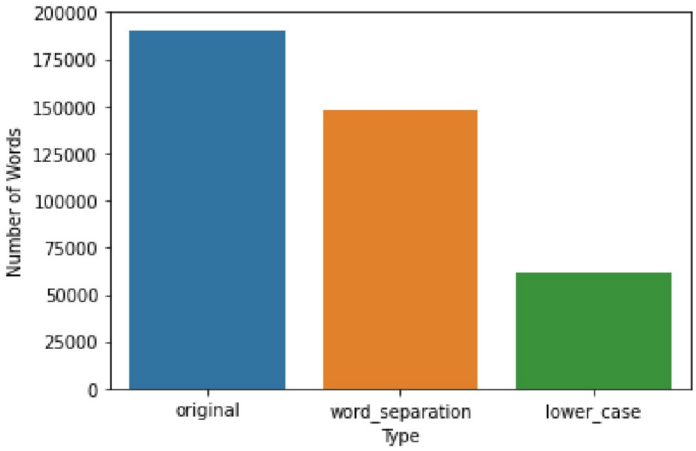

Figure 13: Tamanho do vocabulário após transformação para minúsculas

3.4.3 *POS Tagging*

De seguida, e tendo em conta que a estrutura sintática não está a ser tomada em consideração, podemos remover palavras que não têm qualquer conteúdo importante no contexto da análise de sentimentos, para reduzir o tamanho do vocabulário. Para realização desta tarefa, todas as palavras vão ser identificadas com a sua devida

classe gramatical (*Part-of-speech Tagging*) e, posteriormente, todas as palavras que não são nomes, verbos, advérbios ou adjetivos são descartadas, já que apenas palavras nestas categorias têm conteúdo emocional Martins et al. (2020). Na tabela 12 podemos observar um exemplo do processo de POS *Tagging*, seguido de uma remoção de todas as palavras das categorias mencionadas anteriormente. Depois, na figura 14 podemos observar a redução do tamanho do vocabulário após a execução deste processo.

Opinião	*[“a”, “very”, “fine”, “tuned”, “course”, “used”, “as”, “a”, “warm”, “up”, “course”, “for”, “deep”, “learning”, “highly”, “recommended”]*
POS Tagging	*[(’a’, ’DT’), (’very’, ’RB’), (’fine’, ’JJ’), (’tuned’, ’VBN’), (’course’, ’NN’), (’used’, ’VBN’), (’as’, ’IN’), (’a’, ’DT’), (’warm’, ’JJ’), (’up’, ’RP’), (’course’, ’NN’), (’for’, ’IN’), (’deep’, ’JJ’), (’learning’, ’NN’), (’highly’, ’RB’), (’recommended’, ’VBD’)]*
Remoção de palavras dispensáveis	*[“very”, “fine”, “tuned”, “course”, “used”, “warm”, “up”, “course”, “deep”, “learning”, “highly”, “recommended”]*

Table 12: Processo de POS *Tagging* e remoção de palavras

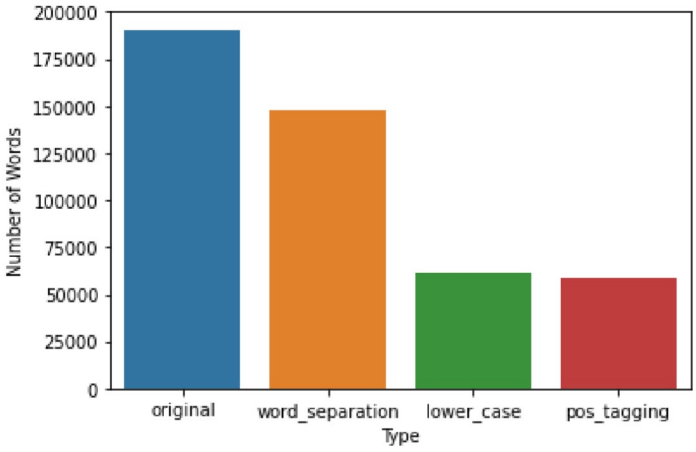

Figure 14: Tamanho do vocabulário após POS *Tagging* e remoção de palavras

Na realidade, este não é dos processos mais comuns para remover palavras que não têm conteúdo emocional, sendo que o mais comum é o processo de remoção de *stopwords*. Em primeiro lugar, o processo de remoção de *stopwords* é mais simples de implementar e, em segundo lugar, é menos custoso em termos computacionais. Ainda assim, o processo descrito anteriormente foi escolhido por duas razões:

1. No processo de remoção de *stopwords*, todas as palavras presentes numa lista pré-definida são eliminadas do vocabulário e, assim, existe um controlo menor sobre quais as palavras que serão eliminadas. Utilizando POS *Tagging*, existe um controlo total sobre quais as categorias de palavras que serão descartadas. Uma

forma de contrariar isto, utilizando na mesma a remoção de *stopwords*, é através da criação manual da lista de palavras descartáveis, o que acaba por ser uma tarefa muito custosa em termos de tempo e recursos.

2. O processo de POS *Tagging* será provado extremamente valioso na fase de identificação de padrões sintáticos, como será demonstrado mais à frente. Por essa mesma razão, é mais eficiente fazê-lo já nesta fase com o propósito de remover as palavras não necessárias à análise de sentimentos, e reutilizá-lo na fase de deteção dos padrões.

3.4.4 *Lemmatization*

Finalmente, a última fase do processo de transformação de dados foi a aplicação de *lemmatization* no conjunto de dados. *Lemmatization* consiste em transformar diferentes formas de uma palavra numa forma base. Na realidade é um processo extremamente semelhante ao processo de *stemming* que já foi explicado anteriormente, mas revela-se um pouco mais sofisticado, já que, ao contrário de apenas transformar uma palavra na sua respetiva palavra raiz, usa o contexto das palavras para as ligar a outras palavras semelhantes em termos de significado. A tabela 13 mostra alguns exemplos de *lemmatization* aplicado sobre a algumas palavras. O resultado da aplicação de *lemmatization* a uma opinião pode ser visto de seguida na Tabela 14.

Palavra original	***Lemma***
rocks	*rock*
better	*good*

Table 13: Exemplos de *lemmatization*

Opinião	***Lemmatization***
[“very”, “fine”, “tuned”, “course”, “used”, “warm”, “up”, “course”, “deep”, “learning”, “highly”, “recommended”]	*[“very”, “fine”, “tune”, “course”, “use”, “warm”, “up”, “course”, “deep”, “learn”, “highly”, “recommend”]*

Table 14: Exemplo de aplicação de *lemmatization* a uma opinião

E, finalmente, a figura 15, mostra-nos o tamanho do vocabulário após a aplicação de *lemmatization* ao conjunto dados.

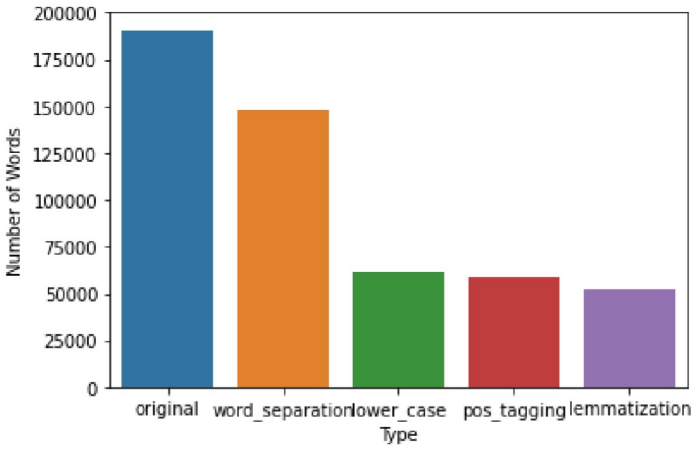

Figure 15: Tamanho do vocabulário após aplicação de *lemmatization*

3.4.5 *Pré-processamento - Resultados e Conclusões*

Após a fase de pré-processamento, foi feita uma pequena análise dos dados como forma de verificação da integridade e da qualidade destes. Como podemos verificar, observando a figura 16, existem algumas palavras que ocorrem com bastante frequência no conjunto de dados e que não revelam qualquer tipo de importância. Estas palavras são nomeadamente palavras com apenas uma letra como "i", "s" e "t". Sendo assim, foi feita uma remoção de todas as palavras com apenas uma letra que estivessem presentes no vocabulário.

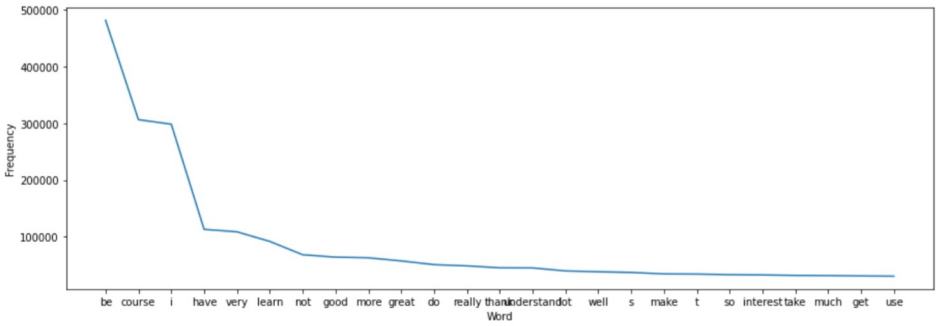

Figure 16: Número de ocorrências das 25 palavras mais comuns após pré-processamento

Em jeito de conclusão, a fase de pré-processamento permitiu realizar uma análise e aplicar um conjunto de transformações ao conjunto de dados recolhido. O processo de transformação dos dados foi dividido em diferentes etapas, sendo que cada etapa foi escolhida tendo por base, principalmente, a redução do tamanho do vocabulário dos dados que, consequentemente, reduz a complexidade do sistema em geral. Se revermos a figura 15 e a tabela 5, podemos verificar que o vocabulário original continha 190493 palavras e, após a aplicação do pré-processamento, ficou com cerca de 60000 palavras, o que corresponde, aproximadamente, a uma redução

de 70% no tamanho do vocabulário. De notar, claro, que este conjunto de transformações não é aconselhado a todos os sistemas de análise de sentimentos, especialmente sistemas que utilizem a estrutura sintática do texto como característica importante na análise de sentimentos. Para além disso, é sempre importante ter em conta os modelos que vão ser utilizados na fase de classificação, já que modelos diferentes podem obter resultados melhores com um conjunto de dados de treino diferente, logo o pré-processamento pode, e deve, ser alterado tendo em conta os resultados dos modelos.

4

CLASSIFICAÇÃO DE SENTIMENTOS

4.1 O PROCESSO DE CLASSIFICAÇÃO

De acordo com o que já foi analisado no capítulo 2, podemos optar pela utilização de vários tipos de modelos, sejam eles baseados em aprendizagem supervisionada, não supervisionada ou até mesmo híbrida. Para que possamos tomar a decisão mais acertada, podemos e devemos utilizar o resultado do processo de preparação de dados como base. Sendo assim, e tendo em conta as características dos dados resultantes do pré-processamento, estruturámos o processo de classificação inicial da forma como está ilustrado na Figura 17:

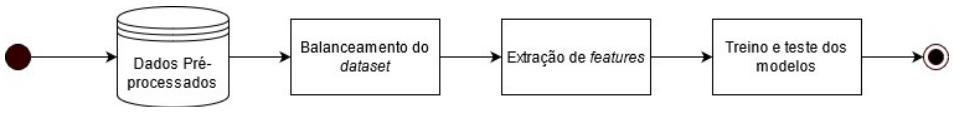

Figure 17: Processo de classificação inicial

Como já foi referido anteriormente, o conjunto de dados processado tem um claro problema de balanceamento. As opiniões com uma classificação de 4 ou 5 são significativamente superiores às restantes opiniões e, por esse mesmo motivo, é necessário, em primeiro lugar, balancear o conjunto de dados para que os modelos não se tornem tendenciosos. De seguida, e tendo em conta que os dados utilizados não contêm qualquer tipo de estrutura sintática, devemos recorrer a métodos de extração de atributos que sejam adaptáveis aos modelos que vão ser utilizados. Finalmente, tendo em conta que os dados extraídos contêm uma coluna representativa da classificação destes, possibilitando assim a aprendizagem supervisionada que tem vindo a obter ótimos resultados para processos de classificação de sentimentos, serão treinados e testados modelos baseados em aprendizagem supervisionada.

Como será possível observar nas secções seguintes, este processo revelou-se insuficiente. A necessidade de classificar as opiniões num valor entre 1 e 5 aumentou significativamente a complexidade do problema, o que resultou numa performance significativamente abaixo do esperado. Por essa mesma razão, foi necessário desenvolver um método que permitisse aumentar a eficácia dos modelos, tendo assim surgido a ideia de utilizar a existência de padrões sintáticos em opiniões como um atributo extra, algo que será explicado extensivamente na secção 4.4. Assim, o processo anteriormente apresentado na figura 17 foi modificado, ficando com a configuração que a Figura 18 ilustra:

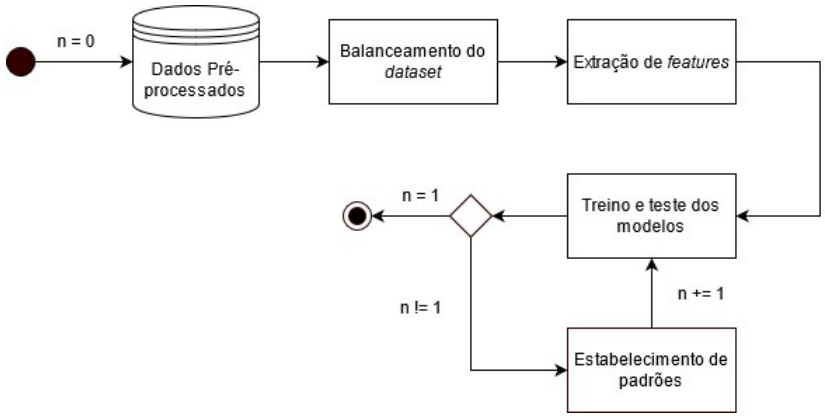

Figure 18: Processo de classificação final

A grande diferença entre o processo inicial e o final é, de facto, a utilização de padrões sintáticos, que foram utilizados para melhorar os modelos. Inicialmente é feito o treino e o teste dos modelos utilizando apenas o processo da figura 17, mas é depois feita uma iteração onde são estabelecidos os padrões sintáticos para as diferentes opiniões e, de seguida, são treinados e testados novamente os modelos.

Nas secções seguintes este processo será detalhado um pouco mais, apresentando-se, então, as técnicas e os modelos desenvolvidos bem como a análise dos resultados dos variados modelos.

4.2 TÉCNICAS E MODELOS DESENVOLVIDOS

Para cada uma das fases do processo de classificação de sentimentos existem inúmeras técnicas ou classificadores que podem ser utilizados. Na realidade, tanto na fase de extração de atributos como na fase de teste e treino dos classificadores devem ser utilizadas mais do que uma técnica, permitindo a comparação entre os diferentes modelos desenvolvidos, garantido assim uma escolha mais ponderada e acertada tendo por base os critérios definidos para avaliação dos modelos. A tabela 15 apresenta os diferentes modelos desenvolvidos numa fase inicial, bem como as técnicas utilizadas nestes:

Técnicas de balanceamento do conjunto de dados	**Modelos de extração de atributos**	**Classificadores**
Undersampling	*Word2Vec*	*Naive Bayes*
		Random Forest
		XG Boost
		Support Vector Machines
	TF-IDF Vectorizer	*Naive Bayes*
		Random Forest
		XG Boost
		Support Vector Machines

Table 15: Técnicas e modelos desenvolvidos para classificação de sentimentos

Como podemos verificar, foram inicialmente desenvolvidos 8 modelos diferentes, o que nos permitirá analisar melhor os resultados de cada um. Destes 8 modelos apenas 2 serão selecionados para uma fase posterior, na qual será feito um estabelecimento de padrões sintáticos de forma a tentar aumentar a eficácia dos classificadores. Esta escolha será baseada na precisão dos modelos, sendo que o objetivo principal desta dissertação é obter a classificação mais precisa possível para qualquer tipo de opinião, seja ela negativa, positiva ou neutra.

4.3 UM PRIMEIRO MODELO DE CLASSIFICAÇÃO

O primeiro passo do modelo de classificação (Figura 17) consiste em balancear o conjunto de dados. No que diz respeito a este processo temos várias opções: *oversampling*, *undersampling* ou uma combinação de ambos. *Oversampling* consiste em criar dados artificiais para uma classe com um número inferior de dados utilizando técnicas de replicação dos dados originais, duplicando, triplicando ou mesmo igualando o número de dados de treino da classe de minoria ao número de dados de uma classe de maioria. Já *undersampling* consiste no processo inverso, no qual o objetivo é eliminar dados de uma classe de maioria, aproximando assim a quantidade destes à quantidade de dados de uma classe de minoria. É ainda possível combinar estes dois métodos. Por forma a facilitar a decisão de qual processo utilizar, podemo-nos basear no seguinte: o número de dados da classe cinco excede em 8 vezes o número de dados da classe um. Utilizando *oversampling* para igualar o número de dados de treino da classe um com o da classe cinco, estaríamos a octuplicar o número de dados de treino da classe um, resultando assim, muito possivelmente, numa situação de *over-fitting*. Apesar da utilização de *undersampling* não ser a mais conveniente já que serão perdidas grandes quantidades de dados não só da classe cinco, mas também da classe quatro, é possivelmente a mais sensata, isto porque, mesmo reduzindo o número de dados de treino de todas as classes para o número de dados de treino da classe um, ficaremos ainda com uma quantidade muito significativa de dados para treinar os modelos e com um conjunto de dados balanceado. Sendo assim, foi utilizado *undersampling* aleatório (método simples de *undersampling*) e a distribuição de dados pelas cinco classes no final do processo foi a seguinte:

Classificação	Nº de opiniões
1	12240
2	12240
3	12240
4	12240
5	12240

Table 16: Número de opiniões por classe

Agora que temos um conjunto de dados balanceado, é necessário fazer a sua tradução para um formato que seja facilmente compreendido pelos modelos de aprendizagem automática. É possível alcançar isto através da utilização de técnicas de extração de atributos como é o caso das técnicas "*TF-IDF*" e "*Word2Vec*", que transformam os textos de opinião em vetores. Ambos os modelos foram treinados com o corpus recolhido pelo *web scrapper*, e foram posteriormente aplicados a todos os textos de opinião do conjunto de dados.

Finalmente, tanto os vetores gerados pelo modelo "*TF-IDF*", como os gerados pelo modelo "*Word2Vec*", foram utilizados para treinar e testar modelos de aprendizagem automática. Na realidade, para ambos os casos, os modelos utilizados foram os mesmos. A escolha baseou-se principalmente na recolha de informação que foi realizada na fase de pesquisa sobre o tema da análise de sentimentos. Os modelos escolhidos foram então: "*Multinomial Naive Bayes*", "*Random Forest*", "*XG Boost*" e "*Support Vector Machines*".

4.3.1 *Análise de Resultados Inicial*

Para uma análise de resultados mais generalista e assertiva, utilizámos a técnica de *cross-validation*. Esta técnica tem como objetivo dividir um conjunto de dados em subconjuntos de dados mutuamente exclusivos que depois serão utilizados, uns para treinar os modelos, e outros para os testar. Mais especificamente, foi utilizado o *Stratified K-Fold* com cinco partições, que permite preservar a percentagem de dados de treino e de dados de teste para cada uma das partições. Na tabela 17 podemos ver a precisão registada após a aplicação de *cross-validation* às oito abordagens diferentes:

Modelo de seleção de atributos	**Modelo de aprendizagem automática**	**Precisão**
TF-IDF Vectorizer	*Multinomial Naive Bayes*	0.47
	Random Forest	0.54
	XG Boost	0.5
	Support Vector Machines	0.55
Word2Vec	*Multinomial Naive Bayes*	0.4
	Random Forest	0.52
	XG Boost	0.5
	Support Vector Machines	0.5

Table 17: Resultados dos modelos após aplicação de cross-validation

Como podemos verificar, os modelos que apresentaram melhores resultados foram o *Random Forest* e *Support Vector Machines*, ambos suportados pela utilização do modelo *TF-IDF* na fase de seleção de atributos, com 54% e 55% de precisão respetivamente. Como é óbvio, estes valores não são satisfatórios e, por isso, é necessário investigar o porquê destes resultados.

Tendo em conta a literatura na área da análise de sentimentos, a razão para a ocorrência destes resultados deverá basear-se na dificuldade de classificar textos de opinião em mais do que duas/três classes, já que, usualmente, os problemas de classificação de sentimentos focam-se em classificar textos de opinião como positivos ou negativos, sendo que por vezes é incluída também a classe "neutro". Teoricamente, classificar textos de opinião como negativos ou positivos é uma tarefa mais simples, já que há uma discrepância significativa entre o que é um comentário negativo e um comentário positivo. Por esse mesmo motivo, executamos os mesmos modelos que executamos para uma classificação entre 1 e 5, mas desta vez tendo como alvo de classificação a coluna "polaridade" que contém as classes negativa, positiva e neutra. Podemos observar os resultados dos modelos, após a aplicação de *cross-validation*, na seguinte tabela:

Modelo de seleção de atributos	Modelo de aprendizagem automática	Precisão
TF-IDF Vectorizer	*Multinomial Naive Bayes*	0.66
	Random Forest	0.73
	XG Boost	0.68
	Support Vector Machines	0.73
Word2Vec	*Multinomial Naive Bayes*	0.58
	Random Forest	0.7
	XG Boost	0.68
	Support Vector Machines	0.68

Table 18: Resultados dos modelos para uma classificação ternária (positiva, negativa ou neutra)

Como se pode observar, a eficácia dos modelos aumentou significativamente, atingindo por vezes uma melhoria de quase 20%. Isto indicia que, de facto, os modelos que estão a classificar os textos de opinião entre 1 e 5 estão a ter dificuldade em distinguir os textos avaliados em 1 com os textos avaliados em 2, e os textos avaliados em 4 com os textos avaliados em 5. Para confirmar esta tendência, vamos observar a matriz de confusão dos modelos "*TF-IDF Vectorizer + Random Forest*" (Tabela 19) e "*TF-IDF Vectorizer + Support Vector Machines*" (Tabela 20) para uma classificação entre 1 e 5 (a coluna do lado esquerdo representa a classificação real das opiniões e a coluna em cima representa a classificação dada pelos modelos).

	1	**2**	**3**	**4**	**5**
1	10215	784	802	214	225
2	4287	2427	3849	1028	649
3	1669	1224	6381	2012	954
4	680	463	2471	4528	4098
5	316	139	542	1963	9280

Table 19: Matriz de confusão do modelo "*TF-IDF + Random Forest*"

	1	**2**	**3**	**4**	**5**
1	9610	1393	848	228	161
2	3871	3409	3849	802	309
3	1273	1897	6367	2065	638
4	387	533	2553	5085	3682
5	184	128	473	2221	9234

Table 20: Matriz de confusão do modelo "*TF-IDF + SVM*"

Como era expectável, a classificação dos modelos para as classes 2 e 4 influencia negativamente a sua eficácia geral, já que a precisão para estas classes está compreendida entre os valores de 30 e 40 por cento. Para além disso, podemos verificar que existe alguma dificuldade em distinguir opiniões com uma classificação de 3, de opiniões com uma classificação de 2 ou de 4.

Tendo em conta todos os resultados apresentados nesta secção, podemos depreender que os modelos clássicos de classificação de sentimentos têm alguma dificuldade na obtenção de resultados positivos para a tarefa em causa. Sendo assim, é necessário criar condições para que os modelos se comportem da forma pretendida para esta dissertação, algo que foi estudado e implementado, e que será apresentado nas secções seguintes.

4.4 UTILIZAÇÃO DE PADRÕES SINTÁTICOS

Após analisar os resultados dos primeiros modelos foi possível depreender que a sua eficácia não era, de todo, a pretendida. Sendo assim, foi necessário fazer alguns ajustes que permitissem disponibilizar mais informação relevante aos modelos.

Em 1992, Hearst (1992) desenvolveu um método de aquisição automática de hipónimos em textos de larga escala através do estabelecimento de padrões sintáticos que usualmente revelam a presença de um hipónimo numa frase. Um exemplo de um padrão estabelecido, bem como uma frase onde este padrão está presente, é apresentado abaixo:

- *such NP as NP, * (or | and) NP*
  - "*... works by such authors as Herrick, Goldsmith, and Shakespeare.*"
    - Hipónimo: ("*author*", "*Herrick*")
    - Hipónimo: ("*author*", "*Goldsmith*")
    - Hipónimo: ("*author*", "*Shakespeare*")

Tal como é possível prever a presença de um hipónimo numa frase tendo por base a presença de padrões sintáticos na sua estrutura, talvez também fosse possível prever a classificação de uma frase da mesma forma, isto é, em frases negativas poderão ocorrer mais comumente, ou mesmo exclusivamente, padrões que não ocorrem em frases com uma polaridade neutra ou positiva. Assim, surgiu a ideia de utilizar o método de "*POS Tagging*", que já foi aplicado anteriormente na fase de processamento dos dados, para a identificação destes padrões. O método desenvolvido foi o seguinte:

1. Revertendo à fase de pré-processamento, antes da remoção de palavras não relevantes e do processo de *lemmatization*, todas as palavras foram identificadas tendo em conta a sua classe gramatical. Isto significa que cada frase ficou traduzida a um conjunto de *tags* que representam a classe gramatical de todos as palavras presentes na frase em questão. Esta informação foi guardada no *dataframe* para ser utilizada na fase de treino dos algoritmos.

2. De seguida, o conjunto de dados foi partido em dois sub-conjuntos de dados, um de treino e outro de teste. Apenas o conjunto de dados de treino foi utilizado na identificação de padrões sintáticos para que o modelo não fique enviesado quando receber como *input* o conjunto de dados de teste.

3. No conjunto de dados de treino foi então utilizada a coluna do *dataframe* que continha a informação sobre a classe gramatical das palavras pertencentes às opiniões. Depois, retirámos as sequências de cinco, seis e sete classes gramaticais consecutivas, sendo estas separadas tendo em conta a classificação da opinião respetiva.

4. Logo de seguida guardamos apenas alguns padrões para cada classe. Esta seleção foi feita através do seguinte critério: se para a classe X, a matriz de confusão demonstrasse uma grande quantidade de dados incorretamente anotados como classe Y ou Z, então seriam retirados dos padrões da classe x todos os padrões que também ocorram na classe Y ou na classe Z. Tal como foi possível observar na secção 4.3.1, muitas opiniões com uma classificação de dois foram anotadas como classe um ou classe três no primeiro modelo, logo todos os padrões que ocorram em opiniões de classificação dois e que ocorram também em opiniões de classificação um ou três, não são considerados padrões de classe dois.

5. Após a identificação dos padrões foram criadas cinco colunas extra aos vetores criados pelo modelo "*TF-IDF*". Cada coluna teve um valor de um ou zero, sendo que o valor um significa que um padrão característico de uma opinião com classificação x está presente na frase e o valor zero significa que não existe. Esta escolha foi feita tendo em conta que os valores resultantes do modelo "*TF-IDF*" são valores compreendidos entre zero e um. Desta forma o modelo terá mais facilidade em depreender estes valores.

6. Finalmente foram preenchidas estas colunas, tanto nos dados de treino como nos de teste. É importante voltar a referir que os padrões foram identificados apenas tendo por base os dados de treino, logo, não é possível saber se estes padrões são identificáveis nos dados de teste.

Podemos observar na tabela abaixo alguns exemplos de padrões identificados, bem como a classe de opiniões a que pertencem, o número de vezes que ocorrem e algumas das frases em que se verifiquem a ocorrência destes [1]:

Padrão	**Classe**	**Nº de Ocorrências**	**Frases**
VB-NN-RB-RB-IN	2	14	"the instructional videos for this class **move way too quickly for** a beginner..."; "...easy to follow examples that do **make sense however then in** the activities they give you wildly more difficult..."
VBN-CD-IN-DT-JJS	5	18	"...but it turns out to have **been one of the best** classes i have ever taken"; ".. this has **been one of the best** sustainability courses i am taking"

Table 21: Padrões identificados e as suas respetivas classes, nº ocorrências e exemplos de frases.

1 VB – verbo na forma base, NN – nome singular, RB – advérbio, IN – preposição ou conjunção subordinada, VBN – verbo no passado, CD – número cardinal, DT – determinante, JJS – adjetivo superlativo

Se analisarmos a informação da tabela 21, tendo em conta que os exemplos foram selecionados aleatoriamente, podemos retirar alguns indicadores interessantes:

- o padrão "VBN-CD-IN-DT-JJS" traduz-se precisamente no mesmo conjunto de palavras em ambos os exemplos, sendo que ambos são claramente representativos de uma reação positiva a um curso realizado pelos autores da opinião, justificando assim a classificação de cinco estrelas;
- no segundo exemplo para o padrão "VB-NN-RB-RB-IN" podemos depreender que o utilizador gostou de um atributo do curso, mas que, ainda assim, desgostou claramente de um outro atributo, o que pode realmente justificar a classificação de duas estrelas.

Assim, estes padrões mostram potencial para serem uma ferramenta coerente no que diz respeito ao apoio que dão aos modelos de classificação. Para que seja possível confirmar esta suposição, é necessário fazer uma nova análise de resultados, tendo agora em conta a utilização dos padrões sintáticos e a sua influência na eficácia dos modelos.

4.5 ANÁLISE DE RESULTADOS

Após a aplicação dos padrões sintáticos recolhidos para melhoria da eficácia dos modelos executados inicialmente, foi necessário averiguar se, realmente, houve a melhoria pretendida. Seguindo a mesma metodologia que foi utilizada na subsecção 4.3.1, vamos primeiro observar a precisão dos modelos executados e, de seguida, as matrizes de confusão destes.

No que diz respeito ao conjunto de dados original, como já foi referido anteriormente, este foi dividido em conjunto de dados de treino e conjunto de dados de teste e representam 80% e 20% do conjunto de dados original, respetivamente. Para ambos os conjuntos de dados, foi aplicada a verificação de existência de padrões retirados do conjunto de dados de treino.

Já no que diz respeito aos modelos, foram selecionados os dois modelos que obtiveram melhores resultados na análise de resultados inicial: "*TF-IDF Vectorizer + Random Forest*" e "*TF-IDF Vectorizer + SVM*". A ambos foi aplicado o algoritmo "*GridSearchCV*" para otimização dos híper-parâmetros e, após a obtenção destes, executaram-se os modelos para o conjunto de dados de teste. Na Tabela 22 podemos ver os resultados obtidos para as previsões dos modelos aplicados sobre os dados de teste.

Modelo	**Precisão**
TF-IDF Vectorizer + Random Forest	0.69
TF-IDF Vectorizer + SVM	0.69

Table 22: Resultado dos modelos para uma classificação entre 1 e 5 após aplicação de padrões sintáticos.

Assim, podemos verificar que, houve uma melhoria significativa da precisão dos modelos, tendo esta aumentado em cerca de 15% para ambos os modelos. Este resultado é já bastante satisfatório tendo em conta que estamos perante um problema de multi-classificação de conteúdo textual, tarefa que usualmente é de enorme

complexidade no espectro da análise de sentimentos. Olhemos de seguida para a matriz de confusão dos dois modelos que está apresentada na Tabela 23 e na Tabela 24.

	1	**2**	**3**	**4**	**5**
1	2181	170	130	35	31
2	266	1756	272	86	54
3	221	226	1436	345	154
4	114	77	399	1195	586
5	44	19	83	391	1933

Table 23: Matriz de confusão do modelo "*TF-IDF + Random Forest*" após aplicação de padrões sintáticos

	1	**2**	**3**	**4**	**5**
1	2119	232	142	42	12
2	214	1809	313	77	21
3	188	294	1474	378	84
4	77	86	425	1323	460
5	28	28	110	555	1749

Table 24: Matriz de confusão do modelo "*TF-IDF + SVM*" após aplicação de padrões sintáticos

É notável ver a melhoria no conflito que havia nas classes 2 e 4, especialmente na classe 2. Podemos observar que, agora, o modelo consegue diferenciar claramente o que é uma opinião com uma classificação de 2 estrelas de uma opinião com uma classificação de 1 ou de 3, algo que não era possível dizer anteriormente.

Para uma análise ainda mais aprofundada, vamos observar a precisão dos modelos quando confrontados com um problema de classificação ternária (positiva, negativa ou neutra). Para isto foi realizada também uma identificação de padrões sintáticos para as classes "neutra" e "negativa" que, ao contrário do que se passou com a classe "positiva", não foram distinguidas corretamente. Estes foram os resultados:

Modelo	**Precisão**
TF-IDF Vectorizer + Random Forest	0.80
TF-IDF Vectorizer + SVM	0.81

Table 25: Resultado dos modelos para uma classificação ternária após aplicação de padrões sintáticos

À semelhança do que aconteceu para o problema de multi-classificação, houve um aumento significativo da precisão dos modelos para uma classificação ternária. Estes resultados permitem-nos comparar a abordagem seguida nesta dissertação com as abordagens mais comuns da literatura, concluindo assim que a precisão dos modelos executados equivale, podendo até por vezes superar, a precisão dos modelos utilizados frequentemente em tarefas de análise de sentimentos, especialmente quando aplicados a problemas de classificação ternária.

4.6 ÍNDICES PARA A CATEGORIZAÇÃO DE SENTIMENTOS

Para que seja percetível a utilidade dos conceitos desenvolvidos durante este trabalho de dissertação num contexto mais realista, esta secção descreverá o processo de categorização dos sentimentos de um utilizador para com uma plataforma educacional. A plataforma educacional em causa contém uma funcionalidade de resposta a inquéritos para recolher as opiniões dos utilizadores, que abrange múltiplos domínios: inquéritos referentes ao sistema em geral, inquéritos referentes a sessões e inquéritos referentes a questões. Cada inquérito contém um conjunto específico de perguntas que permitem ao utilizador expressar o seu sentimento para com características específicas do sistema, de uma sessão ou de uma questão (e.g. a dificuldade de uma sessão tendo em conta o nível de conhecimento necessário). As respostas a estes inquéritos foram armazenadas em documentos *json*, que estão estruturados de acordo com o domínio do inquérito. A figura 19 revela a estrutura de um documento referente a uma resposta a um inquérito sobre uma questão:


```
{
    "idquestion": "UKEINSBD0156",
    "language": "uk",
    "domain": {
        "study_cycle": "University Course",
        "scholarity": "Informatics Engineering",
        "description": "Database Systems"
    },
    "subdomain": "SQL",
    "subsubdomain": "",
    "difficulty_level": 2,
    "user": "E0045",
    "datetime": "2019-11-11 10:23",
    "opinion": {
        "difficulty": 4,
        "time": 3,
        "clarity": 1,
        "content": 4,
        "knowledge": 2,
        "interactivity": 2,
        "comments": "The text of the question is a little bit confusing."
    }
}
```


Figure 19: Documento de resposta a um inquérito sobre uma questão

Na realidade, este documento contém a informação necessária para derivar um índice que categorize os sentimentos do utilizador, mas necessita de algumas adaptações. Em primeiro lugar foi necessário associar um valor quantitativo ao comentário e em segundo lugar é necessário reestruturar o documento para facilitar a visualização do índice tendo em conta diferentes dimensões, como o domínio da pergunta, a data em que o inquérito foi respondido ou mesmo o utilizador que respondeu ao inquérito. Como tal, foi necessário recorrer à execução de um processo ETL – *Extract, Transform, Load*. Para este caso em específico: a extração serviu para recolher informação relevante dos documentos resultantes dos inquéritos; a transformação atribuirá um valor quantitativo ao comentário; e a fase de carregamento dos dados armazenará os dados num *data warehouse*, algo que de seguida permitirá analisar, manipular e visualizar os dados segundo múltiplas perspetivas (OLAP – *Online Analytical Processing*) Ferreira et al. (2010).

Posto isto, foi necessário definir uma estrutura para o *data warehouse*, tendo em conta a estrutura dos documentos de resposta aos inquéritos e as diferentes perspetivas pelas quais queremos manipular os dados. Para este efeito, foi desenhado um esquema multidimensional utilizando a nomenclatura proposta por Golfarelli et al. (1998) – "Dimensional Fact Model". Neste modelo, um *data warehouse* é representado por esquemas de factos, que são representados por grafos diretos, acíclicos e interligados. Na figura 20 está apresentado o esquema dimensional desenvolvido para os inquéritos sobre questões.

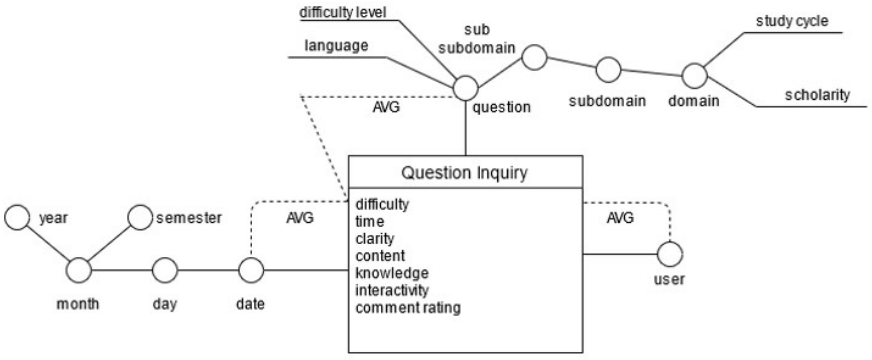

Figure 20: Esquema dimensional desenvolvido para os inquéritos sobre questões

Para facilitar a leitura deste esquema, é importante depreender quais são os componentes de um esquema dimensional e o que os representa no esquema em causa:

- Factos – representam o foco de interesse no processo de decisão (*Question Inquiry*);
- Medidas – atributos com valor, usualmente quantitativos, que descrevem um facto em diferentes pontos de vista (e.g. *difficulty, time* e *clarity*). Para agregar estes valores nas dimensões usámos as suas respetivas médias (*average* - AVG);
- Dimensões – representam atributos discretos que determinam a granularidade mínima adotada para representar factos (*date, user* e *question*);
- Hierarquias – são constituídas por atributos de dimensão discretos interligados, que determinam como é que os factos podem ser agregados para facilitar o processo de decisão. A raiz de uma hierarquia é uma dimensão (e.g. *question*). Os restantes nodos podem ser atributos de dimensão (e.g. *subsubdomain* e *subdomain*) ou não-atributos de dimensão, que contêm informação adicional sobre um atributo de dimensão (e.g. *difficulty level* e *language*).

Assim, com este esquema, podemos claramente caracterizar um inquérito sobre uma questão, durante um dia em específico, respondido por um utilizador.

Posto isto, e tendo já o modelo dimensional definido, é essencial povoar estas bases de dados. Tendo em conta que ainda não existem respostas aos inquéritos na plataforma, foram recolhidas opiniões de alunos da

Universidade do Minho sobre a plataforma *e-learning*: "*Blackboard*". A recolha destas opiniões foi realizada através da criação de um questionário na plataforma "*Google Forms*" que continha apenas duas questões: uma questão de cariz quantitativo, que pedia aos alunos para avaliar o funcionamento geral da plataforma com um valor entre um e cinco, e uma questão de cariz descritivo, onde os alunos, por palavras, podiam expressar a sua opinião no que diz respeito, também, ao funcionamento da plataforma. Assim, extraímos um pequeno conjunto de dados que contém uma opinião e uma avaliação quantitativa que reflete essa opinião, algo que é percetível pelos nossos modelos atuais de classificação de sentimentos.

Em segundo lugar, e já que as opiniões recolhidas estão em português, foi necessário traduzi-las para que os modelos consigam compreender a informação presente nelas. Para isso foi utilizada a libraria *Python – goslate* que utiliza a ferramenta "*Google Tradutor*" como base para traduzir texto. Este método não é de todo ótimo, mas facilita a implementação e será suficiente para testar os modelos numa fase inicial.

Em terceiro lugar, é necessário aplicar o processo de análise de sentimentos apresentado anteriormente neste capítulo a este novo conjunto de opiniões - a aplicação do pré-processamento servirá para reformatar as opiniões de forma que os modelos as compreendam e a aplicação dos modelos de classificação já treinados servirá para fazer a previsão do valor quantitativo representativo dessas opiniões. Os valores quantitativos recolhidos no questionário, serão utilizados para verificação da qualidade dos modelos.

Finalmente, e para criar um ambiente semelhante aquele que será encontrado quando existirem respostas reais aos inquéritos de qualidade da plataforma educacional, serão gerados documentos artificiais (Figura 21) que estarão estruturados de acordo com o modelo multidimensional de dados apresentado na figura 20.


```
{
    "_id": "X6kqGdRzzJhEWewGEHpi7k",
    "user": "Noah Sá",
    "date": {
        "date": "2019-08-08:22:22:11",
        "weekday": "Thursday",
        "month": "August",
        "year": 2019
    },
    "question": {
        "id_question": "BxuoFZEYe7LH77dGLn5QqN",
        "language": "UK",
        "difficulty_level": 3,
        "domain": "Database Systems",
        "study_cycle": "University Course",
        "scholarity": "Informatics Engineering",
        "subdomain": "SQL",
        "sub_subdomain": "queries"
    },
    "difficulty": 2,
    "time": 3,
    "clarity": 3,
    "content": 4,
    "knowledge": 5,
    "interactivity": 3,
    "comment_rating": 1
}
```


Figure 21: Exemplo de um documento armazenado no *data warehouse*

4.6.1 *Visualização dos Índices*

Após a criação dos documentos que contêm uma avaliação sobre uma determinada característica ou tópico da plataforma, ao nível de diferentes índices, é necessário criar, em primeiro lugar, um índice global que categorize o sentimento geral do utilizador e, em segundo lugar, uma plataforma que permita visualizar e analisar os valores deste. Estes índices, quando incorporados em plataformas de visualização e analisados corretamente, podem-se tornar fontes de informação extremamente relevantes para qualquer tipo de utilizador destas plataformas, acrescentando valor de negócio a estas. Para o caso específico da plataforma desenvolvida neste trabalho de dissertação, estes índices permitem, por exemplo, a um professor compreender se uma disciplina que está a lecionar é bem vista pelos alunos e se estes estão a tirar proveito dela, podendo assim ajustar e melhorar a qualidade, por exemplo, das questões que coloca numa ficha de avaliação dessa disciplina. Doutra perspetiva, os índices resultantes dos inquéritos feitos sobre o sistema podem realçar certos aspetos menos positivos da plataforma, ajudando os gestores desta a realizar ajustes no sistema para que este fique, por exemplo, de mais fácil utilização para os utilizadores.

Frequentemente, os índices globais resultam de um valor entre, por exemplo, 1 e 5, atribuído pelos utilizadores de uma plataforma. Este tipo de informação revela-se pobre, já que não é possível depreender quais os aspetos positivos e os aspetos negativos da plataforma, apenas se esta agradou ou não o utilizador na sua generalidade. Para além disso, a ocorrência de situações em que um utilizador atribui um valor que não corresponde aquilo que está expresso no respetivo comentário torna esta informação, por vezes, enganosa. Ora, para este trabalho de dissertação, esta realidade foi tida em consideração e, assim, o índice global é representativo da média ponderada de todos os índices, incluindo o valor retirado da análise de sentimentos para o comentário, atribuindo assim um peso igual a todos os atributos da plataforma.

Como foi referido anteriormente, a existência de índices globais por si só, não é suficiente. É necessário criar ferramentas de análise e visualização, que permitam retirar algum valor de negócio destes índices. Por essa mesma razão, foi desenvolvida uma plataforma de visualização que permite, por exemplo, observar as flutuações dos índices ao longo do tempo, e também o valor destes para, por exemplo, disciplinas específicas, ou até mesmo para tópicos específicos dentro da disciplina. A *framework* utilizada foi a biblioteca de *Python – dash*.

A plataforma desenvolvida permite analisar e visualizar os índices de acordo com algumas das dimensões definidas no esquema de factos apresentado nesta secção. Os filtros aplicáveis no caso dos inquéritos sobre questões são: domínio da questão, subdomínio da questão, data da resposta e utilizador inquirido. Para além disso, são apresentados, na plataforma, atributos como o índice que categoriza os sentimentos, o número de respostas, a avaliação mais alta dada por um utilizador, bem como a avaliação mais baixa. Sendo esta plataforma um *dashboard* interativo, será possível observar as flutuações dos atributos em tempo real.

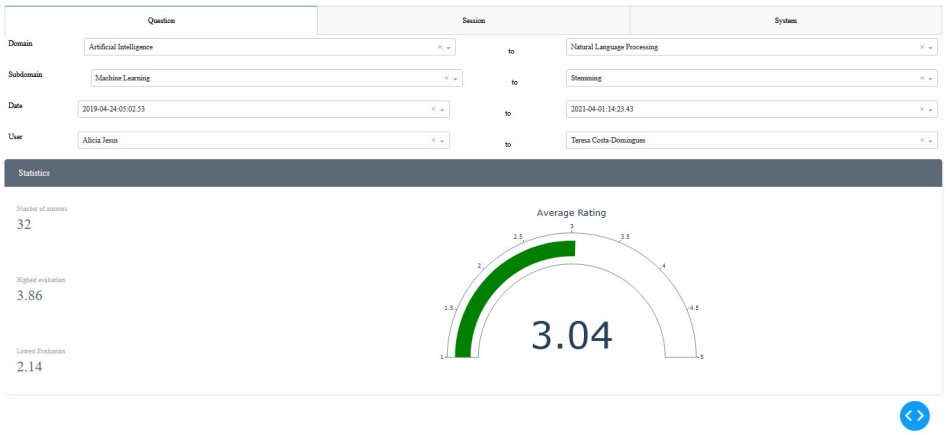

Figure 22: *Dashboard* sem aplicação de filtros

Na Figura 22, podemos observar aquilo que é o dashboard construído sem qualquer aplicação de filtros. Por esse mesmo motivo, podemos inferir que os alunos deram uma cotação média de 3.04 a todas as questões inseridas na plataforma, independentemente da disciplina a que estas questões estão associadas. Podemos também verificar que a cotação mais alta dada por um aluno foi de 3.86 e a cotação mais baixa foi de 2.14. O número total de respostas ao inquérito foi de 32.

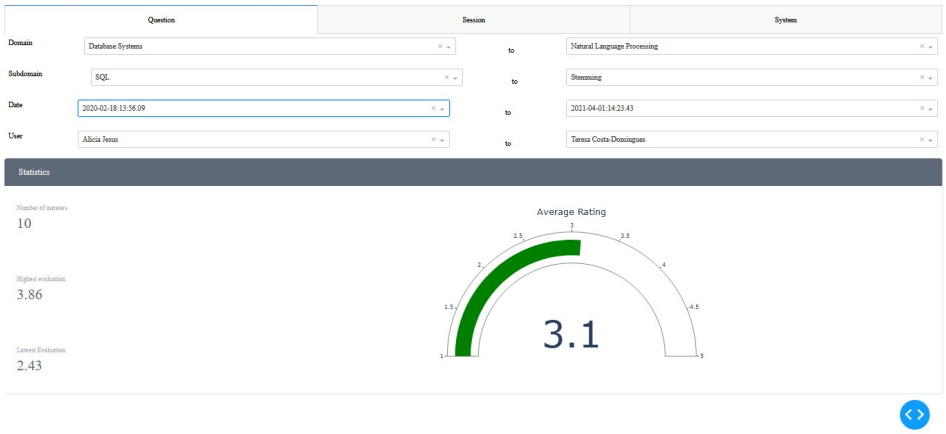

Figure 23: *Dashboard* após aplicação de filtros

Já na Figura 23, podemos observar o dashboard após a aplicação de alguns filtros. Neste caso o intervalo de tempo foi reduzido e o domínio das questões é menos abrangente, apenas contendo questões nas áreas de "Database Systems" e "Natural Language Processing". Os subdomínios foram automaticamente ajustados para subdomínios que estejam inseridos dentro dos domínios selecionados. Podemos verificar que a cotação média dada pelos alunos para as questões filtradas foi de 3.1. Já o número total de respostas foi de 10, com uma cotação máxima de 3.86 e uma mínima de 2.43.

Esta plataforma, apesar de ainda não estar completa, já contém informações que podem ajudar um professor a, por exemplo, compreender se os alunos estão satisfeitos com uma determinada questão. A própria possibilidade de estabelecer um período de tempo para filtrar as respostas é muito importante, já que se podem observar as flutuações do índice. Isto está explícito na transição da Figura 22 para a Figura 23, onde verificamos que entre 2020 e 2021 os alunos estão ligeiramente mais satisfeitos do que entre 2019 e 2021, algo que pode representar uma melhoria na qualidade do serviço.

5

CONCLUSÕES E TRABALHO FUTURO

5.1 CONCLUSÕES

A obtenção de uma solução que permita um computador entender completamente aquilo que um humano pretende expressar, é uma tarefa ainda em progresso. Nos últimos anos tem-se observado uma melhoria drástica neste domínio, especialmente com a introdução dos modelos de aprendizagem automática. A análise de sentimentos, fazendo parte do domínio do processamento de linguagem natural, também tem beneficiado desta realidade e, com isso, criaram-se condições para que, com facilidade, seja possível depreender aquilo que, por exemplo, um cliente sente por um produto ou por um serviço. Ainda assim, é raro encontrar soluções no mercado que apostem neste tipo de sistemas de análise de sentimentos. Este panorama foi identificado com facilidade na fase de revisão da literatura. A razão para a falta dessa aposta está, principalmente, na falta de robustez dos sistemas de análise de sentimentos. De qualquer forma, as próprias entidades que poderiam beneficiar da utilização deste tipo de sistemas, negligenciam, frequentemente, as opiniões textualmente expressas pelos utilizadores, bem como, avaliam a qualidade do seu produto ou serviço através de, maioritariamente, índices estáticos. É, por isso, fundamental melhorar a qualidade destas soluções, não só a nível de eficácia como de performance, para que não seja um risco para as empresas o investimento nelas. Certamente, uma melhoria e uma aposta maior em sistemas de análise de sentimentos resultaria numa melhoria significativa dos produtos ou serviços das entidades, e, consequentemente, uma maior satisfação dos clientes.

No que diz respeito à fase de pré-processamento, esta deve ser extremamente bem planeada e estudada antes de implementada. Tendo em conta que todas as tarefas posteriores ao pré-processamento dos dados dependem fortemente dos seus resultados, um mau planeamento resulta, geralmente, num atraso significativo no andamento do projeto. Um pré-processamento bem executado e ponderado permite compreender a estrutura e a qualidade dos dados na íntegra, o que inevitavelmente também resulta numa seleção mais acertada das técnicas de transformação dos dados, tarefa esta que se comprova de elevada dificuldade devido à existência de um vasto número de possibilidades. Por essa mesma razão, nesta dissertação procurou-se sempre definir um processo que se relacionasse muito proximamente com o objetivo principal, que consiste em obter uma solução de análise de sentimentos para opiniões sobre uma plataforma educacional. Esta realidade é mais notória em processos como o processo de coleção do conjunto de dados, em que foram adquiridos dados que se enquadram perfeitamente no contexto educacional. No restante processo, foram aplicadas transformações que

visaram sempre atingir um formato de dados que fosse facilmente aplicável aos modelos de seleção de atributos já previamente selecionados ("*TF-IDF*" e "*Word2Vec*").

Relativamente à fase de classificação de sentimentos, é importante definir quais as técnicas e os modelos a desenvolver segundo alguns critérios. Em primeiro lugar, é sempre importante ter uma extensa variedade nos modelos de aprendizagem automática e nos modelos de seleção de atributos. Esta variedade permite comparar a precisão dos modelos desenvolvidos, bem como outras métricas como a precisão e o tempo de execução. Isto permite que, posteriormente, se faça uma seleção dos melhores modelos para, por exemplo, aperfeiçoá-los ou mesmo para serem utilizados já num ambiente de produção. Em segundo lugar, e apesar da necessidade de uma variedade grande de modelos para teste, também é importante ser seletivo de acordo com os modelos que demonstraram melhores resultados em trabalhos prévios, logo faz sentido incluir modelos como por exemplo "*Support Vector Machines*" e "*Multinomial Naive Bayes*".

Nesta dissertação procurou-se cumprir ambos os critérios, utilizando dois modelos para seleção de atributos – "*TF-IDF*" e "*Word2Vec*" – e quatro modelos de aprendizagem automática – "*SVM*", "*Multinomial Naive Bayes*", "*XGBoost*" e "*Random Forest*". Numa fase inicial, foi percetível que o modelo "*TF-IDF*" obtinha resultados ligeiramente melhores comparativamente ao modelo "*Word2Vec*" para todos os modelos de aprendizagem automática. Para além disso, os modelos "*SVM*" e "*Random Forest*" também apresentaram melhores resultados do que os modelos "*XGBoost*" e "*Multinomial Naive Bayes*". Assim selecionaram-se as duas combinações "*TF-IDF + Random Forest*" e "*TF-IDF + SVM*" para posteriormente otimizar através da deteção de padrões sintáticos, algo que, logicamente, poupa imensos recursos e tempo já que de oito possíveis combinações sobraram apenas duas para serem otimizadas.

Como já foi mencionado anteriormente, as duas combinações selecionadas não apresentavam os resultados pretendidos. Assim, era fundamental tentar melhorar a eficácia dos modelos para obter modelos minimamente fiáveis. Surgiu assim a ideia inovadora, baseada no trabalho previamente desenvolvido por Hearst (1992), de relacionar padrões sintáticos presentes numa opinião com a sua classificação. De certa forma, esta etapa foi o desafio principal desta dissertação, já que, após longa pesquisa, se pôde concluir que os padrões de Hearst nunca foram utilizados no contexto da análise de sentimentos, algo que faz sentido já que são padrões muito específicos e difíceis de aplicar se não soubermos exatamente aquilo que estamos à procura. Para reduzir a complexidade, utilizou-se "*POS Tagging*" e verificou-se uma melhoria significativa dos modelos, o que indica uma certa tendência para a ocorrência de determinados padrões sintáticos dependendo da polaridade de uma opinião. Este realmente é o fator mais inovador desta dissertação e algo que se pode provar valioso para trabalhos futuros na área da análise de sentimentos.

Finalmente, a etapa de criação de uma ferramenta de visualização de índices que categorizem os sentimentos expressos por utilizadores de uma plataforma educacional provou-se, também, desafiadora. Neste tipo de tarefas é extremamente importante definir qual é a informação que queremos visualizar e como a queremos visualizar. Dessa forma, é possível estabelecer um plano que permite reduzir a quantidade de dados desnecessários e estruturar de uma forma mais clara os dados requeridos para visualização. Esse foi o princípio utilizado para o desenvolvimento deste ponto da dissertação, definindo inicialmente quais as dimensões sobre as quais queremos observar a satisfação dos utilizadores da plataforma. Assim, foi possível estabelecer a estrutura do

data warehouse que armazena os dados úteis e que faculta, de uma forma simples e estruturada, esses dados ao *dashboard*. Por conseguinte, os gestores da plataforma conseguirão analisar a qualidade do seu produto e ajustá-lo, tendo em conta o valor do índice, de acordo com as necessidades dos utilizadores.

5.2 TRABALHO FUTURO

Neste trabalho de dissertação foi possível atingir resultados que consideramos serem de boa qualidade, já que foram obtidos resultados com uma qualidade superior aos resultados da generalidade das soluções de análise de sentimentos, especialmente tendo em conta a dificuldade usual em construir modelos que permitem classificar opiniões num espetro superior áquilo que é a polaridade. Ainda assim, podemos identificar algumas melhorias que podem ser aplicadas:

- Tendo em conta que a previsão feita pelos modelos é algo que deve fazer parte da estrutura da plataforma de apoio ao ensino, seria interessante criar um processo automático que, quando um questionário é respondido por parte de um aluno, inicie automaticamente e permita retirar um valor compreendido entre um e cinco para esse comentário.
- A automatização do processo pode ainda ser mais estendida, criando um mecanismo que não só recolhe as informações relativas à avaliação feita por escrito, mas também dos restantes índices, para depois criar novos documentos que serão guardados numa base de dados e que serão aproveitados pela componente de visualização de índices para apresentar, em tempo real, a satisfação dos estudantes relativamente a um componente específico da plataforma.
- A interface gráfica pode ser melhorada, especialmente em termos de design.
- Neste momento, a classificação dos comentários em português está a ser realizada através da utilização de um tradutor para inglês, já que os modelos estão treinados apenas para dados em inglês, isto devido à falta de recursos a nível de opiniões em português. Como tal, na fase de tradução há sempre uma perda significativa de informação, algo que reduz claramente a qualidade das opiniões, atingindo assim negativamente a eficácia dos modelos. Ainda para mais, a utilização de um tradutor básico, algo que foi utilizado neste trabalho de dissertação, resulta numa perda ainda maior dessa informação. Ora, sendo assim, existem duas opções para melhoria:
  - Utilização de um tradutor de qualidade superior – as redes neuronais transformer têm obtido resultados excelentes em problemas de tradução e seriam, certamente, uma melhoria nesse sentido.
  - A criação ou obtenção de um conjunto de dados em português para treinar os modelos – este tipo de abordagem seria, possivelmente, uma melhoria porque, como a plataforma é focada apenas no contexto educacional e as avaliações serão feitas em português, evitar-se-ia a necessidade de traduzir as opiniões e, consequentemente, a perda de informação.

- Existe ainda alguma incerteza por parte dos modelos em reconhecer a diferença entre opiniões com uma classificação de quatro e opiniões com uma classificação de cinco. Talvez, através da identificação de mais padrões para estas classes, os modelos melhorem o seu desempenho nesse sentido.
- Apesar de ter sido possível concretizar os trabalhos sem qualquer problema e validar os resultados de forma efetiva, seria vantajoso testar os modelos de classificação de sentimentos com um maior número de casos de opiniões gerais dos alunos.

Estes são apenas alguns aspetos que podem ser otimizados, mas, obviamente, existem outros que podem ser desenvolvidos. Foram apresentados vários desafios e dificuldades nos sistemas tradicionais de análise de sentimentos, sendo que a maior parte deles não foram resolvidos na abordagem utilizada durante o desenvolvimento desta dissertação. A identificação de negações e de sarcasmo e o tratamento de frase objetivas e condicionais poderão ser técnicas a implementar futuramente, já que melhorarão, muito possivelmente, a eficácia dos modelos desenvolvidos.

BIBLIOGRAPHY

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

Abd Samad Hasan Basari, Burairah Hussin, I Gede Pramudya Ananta, and Junta Zeniarja. Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. *Procedia Engineering*, 53:453–462, 2013.

Erik Boiy and Marie-Francine Moens. A machine learning approach to sentiment analysis in multilingual web texts. *Information retrieval*, 12(5):526–558, 2009.

Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

Felipe Bravo-Marquez, Marcelo Mendoza, and Barbara Poblete. Meta-level sentiment models for big social data analysis. *Knowledge-based systems*, 69:86–99, 2014.

Wilas Chamlertwat, Pattarasinee Bhattarakosol, Tippakorn Rungkasiri, and Choochart Haruechaiyasak. Discovering consumer insight from twitter via sentiment analysis. *J. Univers. Comput. Sci.*, 18(8):973–992, 2012.

Min Chen, Shiwen Mao, and Yunhao Liu. Big data: A survey. *Mobile networks and applications*, 19(2):171–209, 2014.

Robert A Cummins, Richard Eckersley, Julie Pallant, Jackie Van Vugt, and RoseAnne Misajon. Developing a national index of subjective wellbeing: The australian unity wellbeing index. *Social indicators research*, 64(2): 159–190, 2003.

Erkin Demirtas and Mykola Pechenizkiy. Cross-lingual polarity detection with machine translation. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, pages 1–8, 2013.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dongxu Duan, Weihong Qian, Shimei Pan, Lei Shi, and Chuang Lin. Visa: a visual sentiment analysis system. In *Proceedings of the 5th international symposium on visual information communication and interaction*, pages 22–28, 2012.

Ronen Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

João Ferreira, Miguel Miranda, António Abelha, and José Machado. O processo etl em sistemas data warehouse. In *INForum*, pages 757–765, 2010.

Marc Fleurbaey. Beyond gdp: The quest for a measure of social welfare. *Journal of Economic literature*, 47(4): 1029–75, 2009.

Matteo Golfarelli, Dario Maio, and Stefano Rizzi. The dimensional fact model: A conceptual model for data warehouses. *International Journal of Cooperative Information Systems*, 7(02n03):215–247, 1998.

Zhen Hai, Kuiyu Chang, and Jung-jae Kim. Implicit feature identification via co-occurrence association rule mining. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 393–404. Springer, 2011.

Vasileios Hatzivassiloglou, Judith L Klavans, Melissa L Holcombe, Regina Barzilay, Min-Yen Kan, and Kathleen McKeown. Simfinder: A flexible clustering tool for summarization. 2001.

Marti A Hearst. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*, 1992.

Stefano Maria Iacus, Giuseppe Porro, Silvia Salini, and Elena Siletti. Social networks, happiness and health: from sentiment analysis to a multidimensional indicator of subjective well-being. *arXiv preprint arXiv:1512.01569*, 2015.

Anjali Ganesh Jivani et al. A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl*, 2(6):1930–1938, 2011.

Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

Daekook Kang and Yongtae Park. Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and vikor approach. *Expert Systems with Applications*, 41(4):1041–1050, 2014.

Hanhoon Kang, Seong Joon Yoo, and Dongil Han. Senti-lexicon and improved naïve bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications*, 39(5):6000–6010, 2012.

Subbu Kannan, Vairaprakash Gurusamy, S Vijayarani, J Ilamathi, Ms Nithya, S Kannan, and V Gurusamy. Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, 5(1):7–16, 2014.

Soo-Min Kim and Eduard Hovy. Crystal: Analyzing predictive opinions on the web. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1056–1064, 2007.

Deepika Kumawat and Vinesh Jain. Pos tagging approaches: A comparison. *International Journal of Computer Applications*, 118(6), 2015.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

Yung-Ming Li and Tsung-Ying Li. Deriving market intelligence from microblogs. *Decision Support Systems*, 55(1): 206–217, 2013.

Yung-Ming Li and Ya-Lin Shiu. A diffusion mechanism for social advertising over microblogs. *Decision Support Systems*, 54(1):9–22, 2012.

Isa Maks and Piek Vossen. A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems*, 53(4):680–688, 2012.

Ricardo Martins, José Almeida, Pedro Henriques, and Paulo Novais. Predicting an election's outcome using sentiment analysis. In *World Conference on Information Systems and Technologies*, pages 134–143. Springer, 2020.

Colin D Mathers and Dejan Loncar. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS medicine*, 3(11):e442, 2006.

Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Andrés Montoyo, Patricio Martínez-Barco, and Alexandra Balahur. Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decision Support Systems*, 53(4): 675–679, 2012.

Ramanathan Narayanan, Bing Liu, and Alok Choudhary. Sentiment analysis of conditional sentences. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 180–189, 2009.

Brendan O'Connor, Michel Krieger, and David Ahn. Tweetmotif: Exploratory search and topic summarization for twitter. In *Fourth international AAAI conference on weblogs and social media*, 2010.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*, 2004.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*, 2002.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, pages 19–30, 2016.

Kumar Ravi and Vadlamani Ravi. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems*, 89:14–46, 2015.

Huaxia Rui, Yizao Liu, and Andrew Whinston. Whose and what chatter matters? the effect of tweets on movie sales. *Decision support systems*, 55(4):863–870, 2013.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, 2018.

Joseph E Stiglitz, Amartya Sen, Jean-Paul Fitoussi, et al. Report by the commission on the measurement of economic performance and social progress, 2009.

Songbo Tan, Xueqi Cheng, Yuefen Wang, and Hongbo Xu. Adapting naive bayes to domain adaptation for sentiment analysis. In *European Conference on Information Retrieval*, pages 337–349. Springer, 2009.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. Icwsm—a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *fourth international AAAI conference on weblogs and social media*, 2010.

Peter D Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *arXiv preprint cs/0212032*, 2002.

Jean M Twenge. The age of anxiety? the birth cohort change in anxiety and neuroticism, 1952–1993. *Journal of personality and social psychology*, 79(6):1007, 2000.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

G Vinodhini and RM Chandrasekaran. Sentiment analysis and opinion mining: a survey. *International Journal*, 2 (6):282–292, 2012.

Albert Weichselbraun, Stefan Gindl, and Arno Scharl. Extracting and grounding contextualized sentiment lexicons. *IEEE Intelligent Systems*, 28(2):39–46, 2013.

Rüdiger Wirth and Jochen Hipp. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39. Manchester, 2000.

Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136, 2003.

Jichang Zhao, Li Dong, Junjie Wu, and Ke Xu. Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1528–1531, 2012.