Universidade do Minho

Escola de Engenharia

Departamento de Informática

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

**Ricardo Milhazes Veloso**

Dissertação de Mestrado

2021

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

**Ricardo Milhazes Veloso**

Dissertação apresentada à Universidade do Minho para obtenção do grau de Mestre do Curso de Mestrado Integrado em Engenharia Informática, elaborada sob orientação do Professor Doutor Orlando Manuel de Oliveira Belo.

2021

<dedicatória>

Agradecimentos

(...).

Resumo

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

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 revelação da aplicação 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).*

Abstract

Generation of Well-Being Indexes through Sentiment Analysis

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 exposure of the application 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).*

Índice

1 Introdução 1

1.1 Enquadramento 1

1.2 Motivação e Objetivos 2

1.3 Trabalho Realizado 4

1.4 Estrutura da Dissertação 4

2 Análise de Sentimentos 6

2.1 Contextualização 6

2.2 Definição e Aplicabilidade 7

2.2.1 Nível do documento 8

2.2.2 Nível da frase 8

2.2.3 Nível do aspeto 9

2.2.4 Áreas aplicacionais 11

2.3 O Processo de Análise de Sentimentos 13

2.3.1 Aquisição de Dados 16

2.3.2 Pré-Processamento 18

2.3.3 Classificação de Subjetividade 21

2.3.4 Classificação de Sentimentos 23

2.3.4.1 Aprendizagem supervisionada 24

2.3.4.2 Aprendizagem não supervisionada 27

2.3.4.3 Aprendizagem baseada em dicionários 27

2.4 Desafios e Oportunidades 28

2.5 Sistemas e Utilidade 30

2.5.1 Utilidade 30

2.5.2 Sistemas reais 32

3 Preparação de Dados 34

3.1 O Processo de Preparação de Dados 34

3.2 Definição e Estruturação dos Elementos dos Dados 35

3.3 Análise dos Dados Disponíveis 37

3.4 Tratamento e Transformação dos Dados 40

3.4.1 Separação de palavras e remoção da pontuação 41

3.4.2 *Tokenization* e conversão para letras minúsculas 42

3.4.3 *POS Tagging* 44

3.4.4 *Lemmatization* 46

3.5 Pré-processamento – Resultados e Conclusões 47

4 Classificação de Sentimentos 49

4.1 O Processo de Classificação 49

4.2 Técnicas e Modelos Desenvolvidos 49

4.3 Um Primeiro Modelo de Classificação 49

4.4 Utilização de Padrões Sintáticos 49

4.5 Índices para a Categorização de Sentimentos 49

4.6 Análise de Resultados 49

5 Conclusões e Trabalho Futuro 50

5.1 Conclusões 50

5.2 Trabalho Futuro 50

Referências 52

Índice de Figuras

[Figura 2.1 - Processo de análise de sentimentos para opiniões referentes a produtos 13](#_Toc95753618)

[Figura 2.2 - Processo geral de análise de sentimentos 14](#_Toc95753619)

[Figura 2.3 - Exemplo de aplicação do processo de stemming. 20](#_Toc95753620)

[Figura 2.4 - Modelos de classificação de sentimentos 23](#_Toc95753621)

[Figura 3.1 - Pré-processamento típico 34](#_Toc95753622)

[Figura 3.2 - Recolha de dados 37](#_Toc95753623)

[Figura 3.3 - Número de opiniões por classificação 39](#_Toc95753624)

[Figura 3.4 - Número de ocorrências das 25 palavras mais comuns 39](#_Toc95753625)

[Figura 3.5 - Processo de transformação dos dados 40](#_Toc95753626)

[Figura 3.6 - Tamanho do vocabulário após a separação de palavras 42](#_Toc95753627)

[Figura 3.7 – Tamanho do vocabulário após transformação para minúsculas 44](#_Toc95753628)

[Figura 3.8 - Tamanho do vocabulário após POS Tagging & remoção de palavras 45](#_Toc95753629)

[Figura 3.9 - Tamanho do vocabulário após aplicação de *lemmatization* 47](#_Toc95753630)

[Figura 3.10 – Número de ocorrências das 25 palavras mais comuns após pré-processamento 48](#_Toc95753631)

Índice de Tabelas

[Tabela 2.1 - Exemplo de tokenization de frases 18](#_Toc95753632)

[Tabela 2.2 - Exemplo de remoção de stop-words. 19](#_Toc95753633)

[Tabela 3.1 - Extrato de algumas linhas do *dataset* 38](#_Toc95753634)

[Tabela 3.2 - Dimensões do *dataset* 38](#_Toc95753635)

[Tabela 3.3 - Número de palavras que ocorrem no *dataset* 40](#_Toc95753636)

[Tabela 3.4 – Exemplo de aplicação de separação de palavras 41](#_Toc95753637)

[Tabela 3.5 – Exemplo de aplicação de remoção de pontuação 42](#_Toc95753638)

[Tabela 3.6 - Exemplo de aplicação de um *tokenizer* 43](#_Toc95753639)

[Tabela 3.7 - Exemplo de aplicação de uma transformação para minúsculas 43](#_Toc95753640)

[Tabela 3.8 - Vetor BOW antes da transformação para minúsculas 43](#_Toc95753641)

[Tabela 3.9 - Vetor BOW após a transformação para minúsculas 44](#_Toc95753642)

[Tabela 3.10 – Processo de POS Tagging & remoção de palavras 45](#_Toc95753643)

[Tabela 3.11 - Exemplos de lemmatization 46](#_Toc95753644)

[Tabela 3.12 - Exemplo de aplicação de lemmatization a uma opinião 47](#_Toc95753645)

Capítulo 1

# Introdução

## 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.* [1]). 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 M.* [2], *Cummins et al.* [3]), 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.* [1], *Iacus et al.* [4]).

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, surgem 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.* [4]). 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.* [4]):

* 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.

## 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* [5]), 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 J.* [6]). 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 procurar-se-á analisar 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.

## Trabalho Realizado

## Estrutura da Dissertação

Após o capítulo introdutório, o segundo capítulo descreve toda a pesquisa feita sobre o tema da análise de sentimentos. Inicialmente é feita uma contextualização geral do tema, seguida da apresentação do estado de arte deste, que incluí as diferentes abordagens utilizadas atualmente, a evolução destas nos últimos anos e os desafios atuais.

Os dois capítulos subsequentes ao segundo capítulo explicam todo o processo de desenvolvimento de uma solução que permita calcular e visualizar os índices de bem-estar gerados, sendo que o capítulo de preparação de dados 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 retrata principalmente o método de cálculo dos índices de bem-estar, que, tal como a designação do 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 uma visualização destes, tendo por base um sistema multidimensional de dados.

Finalmente, o capítulo de conclusões e trabalho futuro permite fazer uma retrospetiva daquilo que foi desenvolvido ao longo desta dissertação, analisando criticamente os resultados obtidos e os processos que foram elaborados. Também é feita 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 próximos passos para a integração do sistema no quotidiano.

Capítulo 2

# Análise de Sentimentos

## 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.* [34] 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. 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, 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 se desenvolvem mutuamente, criando assim 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.* [35], podem ser divididos 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**, usualmente é 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**.

## Definição e Aplicabilidade

A análise de sentimentos representa o estudo computacional das opiniões, atitudes ou emoções de um indivíduo relativamente a um determinado assunto, sendo que este assunto pode representar um indivíduo ou conjunto de indivíduos, um evento ou um tópico (*Medhat et al.* [7]). Existem 3 tipos de níveis de análise de sentimentos: ao nível do documento, ao nível da frase e ao nível do aspeto.

### 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 e expressa pelo autor do documento, sobre um assunto em específico. (*Feldman R.* [8])

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.* [9]) 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 P.* [10])

Ainda assim, como sabemos, um documento pode conter múltiplas 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.

### Nível da frase

A esmagadora maioria das implementações de análise de sentimentos ao nível da frase 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 um processo frequente na fase de pré-processamento dos dados. *Yu and Hatzivassiloglou* [11] 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.* [12]) - 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* [13] 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.

Como consideração final, 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.

### 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 o objetivo de reconhecer todas as expressões de sentimento contidas numa opinião e todos os aspetos a que estas se referem. (*Feldman R.* [8])

*Pontiki et al.* [14] 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

* Exemplo 1: “*Este computador tem uma placa gráfica muito boa*”
  + - (1) – “*computador*”
    - (2) – “*placa gráfica*”
    - (3) – positiva

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

O objetivo é identificar as entidades e os respetivos aspetos presentes em cada frase. As estratégias mais utilizadas são:

* **POS (*Part-of-Speech Tagging*)** – 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** – 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.* [15] recorreram a regras de associação para equiparar expressões de sentimento (aspetos implícitos) a aspetos explícitos.

* Exemplo 2: “*Este quadro é pesado*”
  + O aspeto presente nesta opinião é o peso, mas não está mencionado explicitamente.

#### Classificação da polaridade dos aspetos (3):

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.

### 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, é importante focarmo-nos principalmente na classificação de sentimentos.

A classificação de sentimentos pode ser dividida em 4 categorias: 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. (*Ravi and Ravi* [16])

#### 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. A deteção destas características por si só é desafiadora e, por isso, existem atualmente algumas iniciativas para a resolver. *Tsur et al.* [17] criaram um modelo semi-supervisionado para classificar o nível de sarcasmo (valor entre 1 e 5) 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. É, assim, complicado obter um modelo universal eficaz para analisar sentimentos nos diferentes idiomas. Atualmente já existem várias abordagens que visam resolver este problema, sendo que existem duas principais alternativas:

* **Baseada em léxicos** – consiste em classificar uma opinião traduzindo o léxico do idioma original para outro idioma onde já existam recursos que facilitem a classificação. (*Demirtas E.* [18])
* **Baseada em *corpus*** – consiste em criar um *corpus* na linguagem original, devidamente anotado, para treinar um classificador estatístico que faculte a classificação de opiniões. (*Boiy and Moens* [19])

#### 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 onde o classificador vai ser treinado (domínio origem) e o domínio em que 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.* [20] 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.* [21] 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.

## 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 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.* [7], e consiste num processo tipicamente utilizado para analisar sentimentos em opiniões referentes a produtos. A figura 2.1 mostra precisamente uma versão adaptada da implementação descrita por *Medhat et al.* [7]:

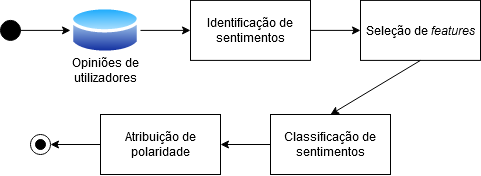


Figura 2.1 - Processo de análise de sentimentos para opiniões referentes a produtos

* A fase de identificação de sentimentos permite a identificação de conteúdo subjetivo e a extração do autor da opinião. Esta tarefa é mencionada por *Ravi and Ravi* [16] 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;
* Já a fase de seleção de *features* 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;
* A atribuição de polaridade consiste em atribuir uma conotação negativa, positiva ou neutra às opiniões, através da cotação das opiniões dada pelo classificador.

Este processo apresentado por *Medhat et al.* [7] 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 são relevantes na mesma.

Uma fase que está subentendida no processo definido por *Medhat el al.* [7] 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* [22]) 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 ricos em informação 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:

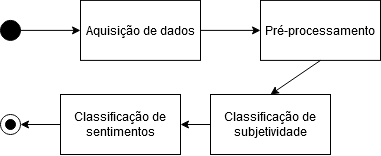


Figura 2.2 - Processo geral de análise de sentimentos

O processo de análise de sentimentos apresentado na figura 2.2, é de facto uma representação que generaliza bastante todos os processos de análise de sentimentos atualmente existentes. Por essa razão, vai ser utilizado como referência para a elaboração do restante 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 todas as imensas e variadas fontes que existem atualmente para recolher informação, com o objetivo principal do processo em si, isto é, 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 e os mais comuns são: *tokenization*, remoção de *stop-words* 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.

### 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 identificar 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* [22], onde podem ser recolhidas 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 os utilizadores se podem expressar livremente sem influências externas e é 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* [13], 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.* [23] 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 scrapers* 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 necessitar 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.* [24] usaram opiniões sobre filmes para testar um sistema híbrido de análise de sentimentos. Estas opiniões foram recolhidas inicialmente do Twitter (blogues e micro blogues) mas, 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 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.

### Pré-Processamento

A fase de pré-processamento é importantíssima para reduzir a complexidade das opiniões e simplificar todo o processo de análise de sentimentos, melhorando a qualidade dos dados iniciais. Tendo em conta que estes dados estarão em formato textual, logo serão representativos de 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.

#### 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, este método não 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, mas, ainda assim, esta abordagem pode ser problemática porque implica a separação de indicadores importantes como *emoticons*. A tabela 2.1 demonstra um exemplo de *tokenization* a uma frase.

|  |  |
| --- | --- |
| Frase | O dia hoje está muito agradável. |
| *Tokens* | [“O”, “dia”, “hoje”, “está”, “muito”, “agradável”, “.”] |

Tabela 2.1 - 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 *stop-words*. Ainda assim, 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.* [25] 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 stop-words

A remoção de *stop-words* 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.* [26]). 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 *stop-words*, sendo que o principal consiste na utilização de dicionários ou listas pré-definidas compostas por *stop-words*,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 2.2 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.

|  |  |
| --- | --- |
| Lista | [“ele”, “ela”, “também”, “até”] |
| Frase | Ele e ela também se dão muito bem. |
| Resultado | e dão muito bem. |

Tabela 2.2 - Exemplo de remoção de stop-words.

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.

#### 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 Ms.* [27]). Esta redução normalmente não afeta a qualidade dos dados já que, é observado frequentemente que palavras morfologicamente parecidas têm as mesmas interpretações semânticas. Ainda assim, 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. Estes métodos são denominados de métodos de truncamento. A figura 2.3 mostra um exemplo geral de aplicação de *stemming*, adaptado do exemplo apresentado por *Kannan et al.* [26]*.*

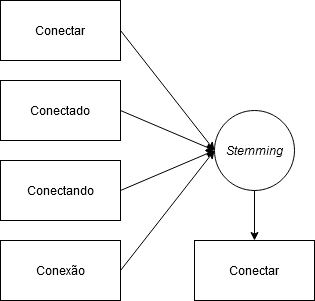


Figura 2.3 - Exemplo de aplicação do processo de stemming.

Relativamente à análise de sentimentos, as vantagens deste processo são similares às vantagens dos dois processos anteriores. Realmente 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 Ms.* [27]).

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*”.

### 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.

* **Conteúdo objetivo** – 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.

Sendo assim, 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. Assim, 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* [28] consideram que a subjetividade pode ser expressa de duas formas:

* **Subjetividade do orador** – consiste na representação dos sentimentos, emoções e perceções do orador.
  + Exemplo 2: “*Ele é um mentiroso*”
    - Representa uma atitude negativa do **orador** para com “*ele*”
* **Subjetividade da personagem** – consiste na representação dos sentimentos, emoções e perceções da personagem.
  + Exemplo 3: “*O ódio dele pela religião*”
    - Representa uma atitude negativa da **personagem** (“*dele*”) para com a religião.
  + Exemplo 4: “*Eles estão entusiasmados pelo serviço*”
    - Representa 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* [29]. 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.* [30] 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 SentiWordnetcriado por *Baccianella et al.* [31], para determinar se uma frase é subjetiva ou não.

### 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 2.4, adaptada de *Medhat el al.* [7], apresenta precisamente esses modelos.

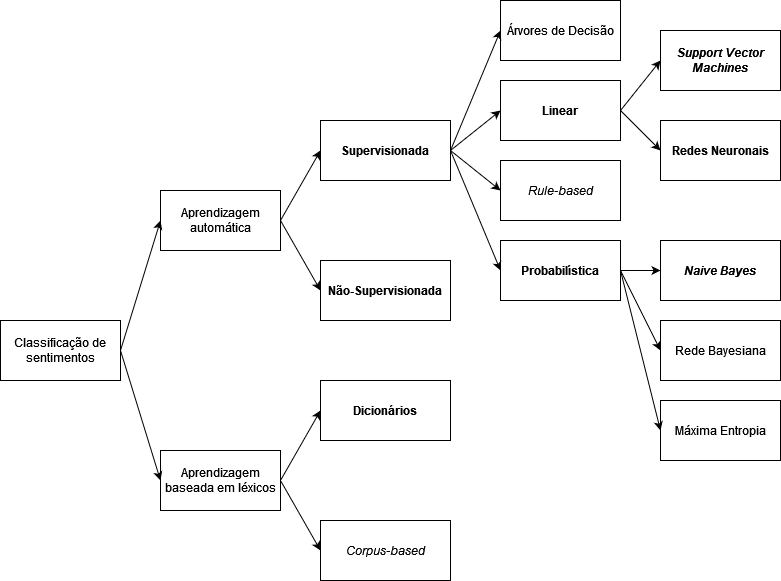


Figura 2.4 - Modelos de classificação de sentimentos

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 são fases importantíssimas para o bom funcionamento dos sistemas de análise de sentimentos, sendo que, para sistemas que implementem aprendizagem supervisionada, ainda mais importantes são. Para treinar estes modelos, é importante recorrer a técnicas de extração e seleção de atributos textuais presentes nos dados já processados. Algumas das técnicas mais comuns são:

* **Presença de termos e a sua frequência** – consiste na 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)** – consiste na identificação de características textuais que indiquem a existência de sentimentos, sendo que também pode ser considerada uma técnica de classificação de subjetividade. Por exemplo, palavras como adjetivos exprimem, tendencialmente, emoções;
* **Palavras de opinião ou frases** – representa 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:
  + Exemplo 5: “*Hoje estou com a telha.*”;
  + Significa que o titular da opinião está de mau humor (polaridade negativa).
* **Negações** – consiste na identificação de palavras negativas que, por norma, altera a orientação da opinião.
  + Exemplo 6: “*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 e/ou frequência de termos em opiniões.

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.* [32] 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* [35] 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.* [32], 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, é lógico que deve ser feita também uma análise bibliográfica destes. Como podemos observar na figura 2.4, os métodos de aprendizagem supervisionada estão partidos em 4 classes principais: classificadores de árvores de decisão, lineares, probabilísticos e baseados em regras. Destes, 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.

#### Multinomial Naive Bayes

O modelo “*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 *dataset*. 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 and Nigam* [36], 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 modelo “*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 *features* utilizadas para treinar o modelo, dividem o espaço de decisão de forma a 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 utilizadas 3 *features* 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 *features* (três) utilizadas no conjunto de dados. Podemos observar um exemplo na figura abaixo:

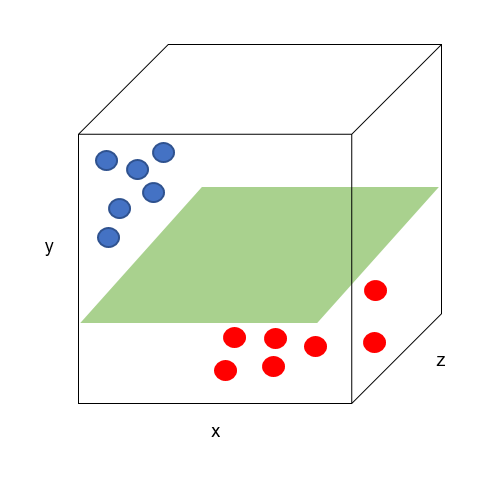


Figura 2.5 - Exemplo de um hiperplano definido para um espaço de decisão tridimensional

Após uma breve introdução ao modelo, é necessário compreender o porquê deste modelo ser, teoricamente, um modelo ótimo para tarefas de classificação de sentimentos. Ora, segundo *Joachims T.* [37], existem algumas evidências teóricas que justificam a boa performance dos modelos SVM quando aplicados a este tipo de tarefas, como por exemplo:

* **a natureza dimensional das *features*** – usualmente quando lidamos com problemas de análise de sentimentos, o número de *features* utilizadas para treinar os modelos costuma ser significativo, algo em concordância com o tamanho dos vocabulários em que os problemas estão inseridos. Sendo que o modelo SVM usa proteção contra *overfitting*, o que não depende do número de *features*, acaba por suportar grandes espaços de decisão;
* **a relevância das *features*** – uma forma de ignorar um grande espaço dimensional de *features*, é considerar que a maior parte delas é irrelevante. Este método retiraria a limitação apresentada anteriormente, mas em problemas que envolvem textos de opinião, retirar *features* pode significar uma perta significativa de informação, tendo já se comprovado que mesmo as *features* 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, que 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:

* 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.* [38]), onde a principal diferença entre estas e as redes neuronais recorrentes é o facto das TNN receberem os dados de uma forma paralelizada.

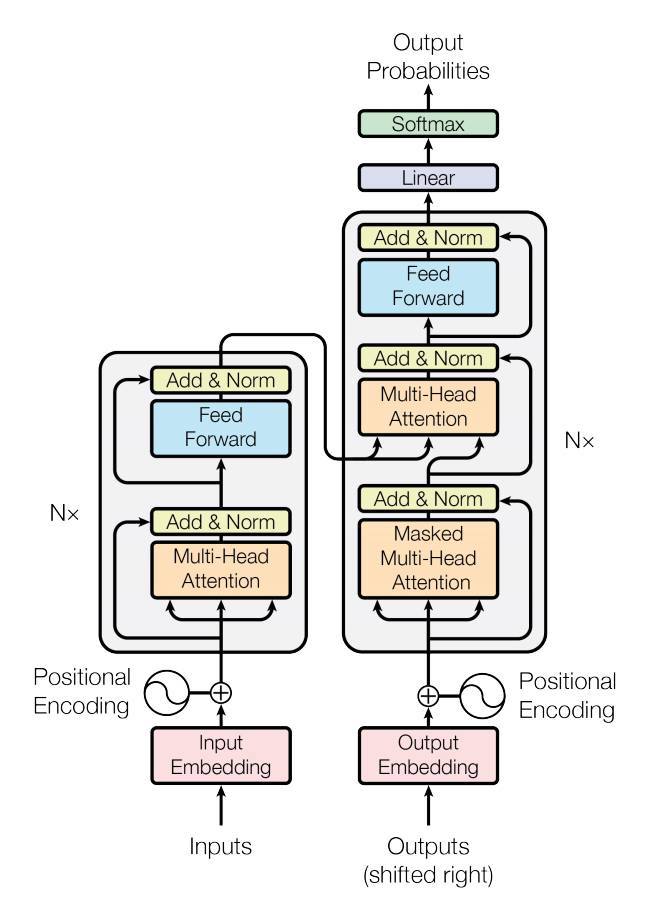


Figura 2.6 - Arquitetura das redes neuronais "*transformer*"

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. O *encoder* recebe então um conjunto de n palavras em português, em simultâneo, e retorna n vetores que representam o significado de cada palavra.

* Inicialmente as palavras 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”*);
* 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”*);
* Estes vetores são então 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 tendo em conta as atenções direcionadas à própria palavra (“*self-attention”*).

Já o decoder recebe o conjunto de palavras traduzidas para inglês

### Aprendizagem não supervisionada

Na análise de sentimentos é frequentemente 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 de treino 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. Podemos então 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.

## 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. Abaixo estão representados os principais desafios referentes à área da análise de sentimentos, bem como algumas implementações de soluções para alguns deles.

* **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;
  + Exemplo 7: “*As sirenes desta ambulância são silenciosas*”
  + Exemplo 8: “*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. Podemos observar esse mesmo efeito no exemplo 6 da secção 2.5.4. 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 raras situações em que conteúdo objetivo expressa sentimentos. Estas situações normalmente identificam-se através da exposição de factos indesejáveis relativamente a uma entidade;
  + Exemplo 9: “*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, onde 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.* [39] elaboraram diferentes métodos para determinar a polaridade de frases condicionais. Em primeiro lugar, estabeleceram diferentes combinações de atributos extraídos dos dados de treino classificados, como por exemplo, palavras de negação e 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.
  + Exemplo 10: “*Se este telemóvel for bom, compro-o*”
* **Sarcasmo** - significa que, quando é expresso um sentimento positivo, em formato textual, na realidade o autor refere-se a um sentimento negativo e vice-versa. 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. (*González-Ibáñez et al. [16]*)
  + Exemplo 11: “*Que jogo bem feito, nem consigo ver a equipa adversária*”

## Sistemas e Utilidade

### Utilidade

Atualmente, a análise de sentimentos pode influenciar positivamente um vasto número de áreas aplicacionais. Na realidade, esta capacidade abrangente não passa por 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 do tema da análise de sentimentos, apresentadas por *Ravi and Ravi* [47], são:

* **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.* [6] 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.* [48] 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* [30] criaram um sistema de marketing inteligente através da observação de opiniões em microblogs. 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* [23] 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 e o Youtube. São considerados uma ferramenta poderosa, já que cria uma experiência 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* [31] 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, e a **previsão de eleições políticas**, onde, 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.

### Sistemas reais

Como podemos depreender, tendo em conta o capítulo anterior, a análise de sentimentos tem uma vasta área aplicacional que, para além de ser vasta, 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:

* **Moodlens -** O Moodlens (*Zhao et al.* [58]) 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, nojento, 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.
* **VISA -** O VISA (*Duan et al.* [13]) 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 todo 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.
* **MSAS -** O MSAS (*Chamlertwat et al.* [8]) (*Micro-blog Sentiment Analysis System*) é um sistema de análise de sentimentos que analisa 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, tais como ecrã, aplicações e câmara.
* **Crystal** - O Crystal (*Kim and Hovy.* [27]), como já foi referido anteriormente, é um sistema de análise de sentimentos que tem como objetivo 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.

Capítulo 3

# Preparação de Dados

## O Processo de Preparação de Dados

Ainda que a aquisição de grandes quantidades de dados seja, atualmente, simples, como foi referido no capítulo anterior, nem sempre estes conjuntos de dados estão estruturados de uma forma desejada. Usualmente verificam-se inconsistências nos dados, como por exemplo a ocorrências de valores nulos e dados desformatados. Por essa mesma razão, é necessária 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**. Podemos observar na figura 3.1 uma representação típica da fase de pré-processamento:



Figura 3.1 - Pré-processamento típico

Na realidade, este esquema é uma versão baseada na metodologia padrão aplicada em processos de mineração de dados – CRISP-DM [59], que será a base para as soluções apresentadas nos capítulos seguintes.

Podemos então concluir que 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 que, por si só, dependem fortemente da fase de preparação de dados.

Nos restantes subcapítulos serão apresentadas todas as fases do processo de preparação de dados, bem como a motivação e explicação por detrás das técnicas utilizadas.

## Definição e Estruturação dos Elementos dos Dados

O primeiro passo do processo de preparação de dados é definir corretamente os 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 para o caso desta dissertação em específico, 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 expresso nesta. Ora, para um melhor entendimento do que estes valores representam em termos de sentimento expresso 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 e o valor 5 representa uma opinião com uma polaridade muito positiva. Sendo 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 porque, por exemplo, 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 melhor 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, serão recolhidas opiniões redigidas em inglês, já que opiniões neste idioma são mais comuns e facilitam substancialmente a fase de pré-processamento.

Após uma pesquisa exaustiva por dados que contenham os requisitos referidos anteriormente, foi possível concluir que, apesar de já existirem vários *datasets* que contêm opiniões anotadas com um valor entre 1 e 5, não existe nenhum *dataset* pré-definido que contenha opiniões de utilizadores sobre plataformas educacionais anotadas com um valor entre 1 e 5. Ainda assim, durante esta pesquisa, foi possível encontrar uma plataforma de cursos online que contém uma vasta quantidade de opiniões sobre os seus cursos/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 3.2 podemos observar o processo geral de recolha de dados:

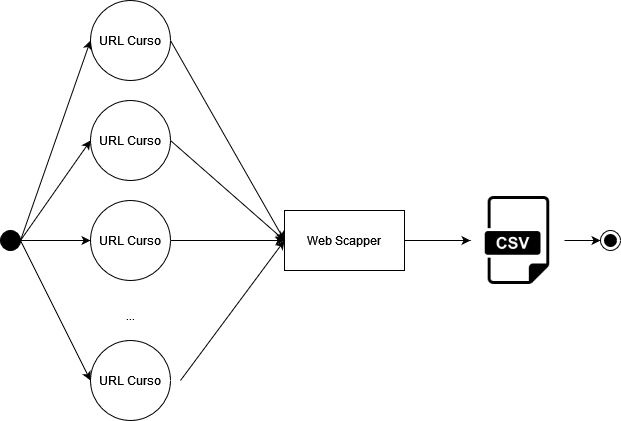


Figura 3.2 - Recolha de dados

Como se pode verificar na figura 3.2, foram recolhidas opiniões de vários cursos da plataforma online em causa, para construir um *dataset* comprido e robusto. Durante o processo de *web scrapping*, é feita uma verificação do idioma da opinião para apenas recolher comentários em inglês.

Para além de todo este processo, foi adicionada uma nova coluna ao *dataset* que contém a descrição da polaridade da opinião, isto é, as opiniões classificadas como 1 ou 2 foram descritas como negativas, as opiniões classificadas como 3 foram descritas como neutras, e as opiniões classificadas como 4 ou 5 foram descritas como positivas. Esta coluna foi apenas adicionada por questões de análise futura.

## 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 *dataset*, que ajudarão a compreender quais os processos de transformação que terão que ser aplicados nas fases posteriores.

|  |  |  |
| --- | --- | --- |
| **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* |

Tabela 3.1 - Extrato de algumas linhas do *dataset*

|  |  |
| --- | --- |
| **Número de linhas** | **Número de colunas** |
| 277287 | 3 |

Tabela 3.2 - Dimensões do *dataset*

Como podemos verificar, olhando para a tabela 3.2, o *dataset* contém 277287 opiniões de utilizadores. Para além disso, cada opinião está associada à sua polaridade e classificação respetiva. O *dataset* não contém valores nulos, o que significa que o *web scrapper* desenvolvido funcionou corretamente.

De seguida é necessário analisar o *dataset* 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.

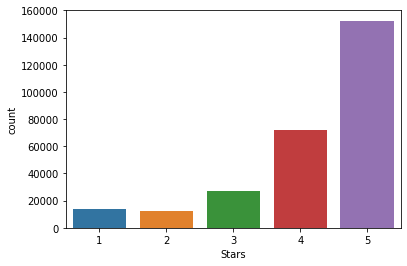


Figura 3.3 - Número de opiniões por classificação

Como podemos verificar, este *dataset* não é exceção, 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 *dataset* na fase de treino e teste dos modelos.

Finalmente, sabendo que a eficácia dos modelos de análise de sentimentos depende muito do vocabulário onde está inserida, este mesmo deve ser analisado.

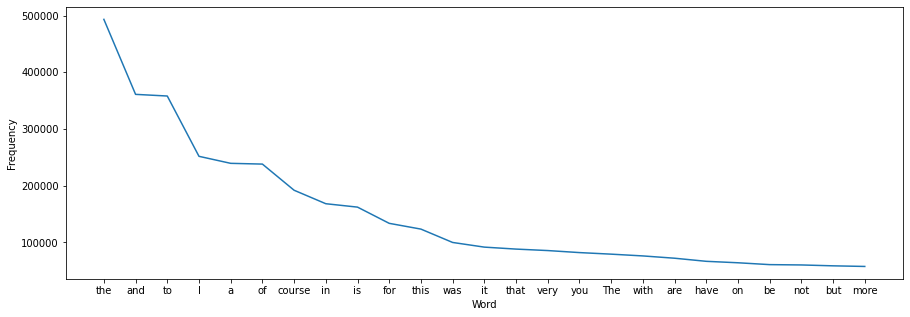


Figura 3.4 - Número de ocorrências das 25 palavras mais comuns

|  |
| --- |
| **Número de palavras** |
| 190493 |

Tabela 3.3 - Número de palavras que ocorrem no *dataset*

Tendo em conta informação da tabela 3.3, é possível inferir que o *dataset* contém 190493 palavras diferentes. Para além disso, se olharmos para a figura 3.4, 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. Isto 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.

## Tratamento e Transformação dos Dados

Depois de analisar extensivamente o conjunto de dados recolhido, é necessário fazer algumas transformações a este. 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 3.5, podemos observar o processo desenhado para a fase de tratamento e transformação de dados, sendo que os capítulos seguintes irão explicar o porquê da sua realização (através da análise de alguns exemplos) e o seu efeito no conjunto de dados final.

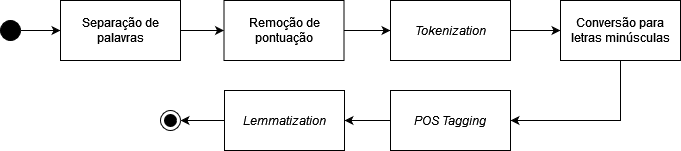


Figura 3.5 - Processo de transformação dos dados

### 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 postadas em plataformas onde não é necessário qualquer tipo de estrutura de escrita, o formato da escrita é usualmente descuidado. Observemos alguns 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 estes comentários estejam incorretamente redigidos 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. A tabela 3.4 demonstra um exemplo da aplicação da separação de palavras a 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* |

Tabela 3.4 – Exemplo de aplicação de separação de palavras

Na figura 3.6, verificamos 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.



Figura 3.6 - Tamanho do vocabulário após a separação de palavras

Após a separação de palavras, que por si só já ajudou a reduzir a complexidade do sistema, será também útil remover a pontuação das opiniões. 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 em que estas opiniões estão inseridas, 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* |

Tabela 3.5 – Exemplo de aplicação de remoção de pontuação

### *Tokenization* e conversão para letras 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 *string* 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 3.6, ilustra 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”*] |

Tabela 3.6 - Exemplo de aplicação de um *tokenizer*

Após o processo de *tokenization*, foi efetuada uma transformação de todas as palavras do vocabulário para palavras minúsculas. A tabela 3.7 ilustra 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”*] |

Tabela 3.7 - Exemplo de aplicação de uma transformação para minúsculas

Mais uma vez, o objetivo deste processo é 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 BOWcomo 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:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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 |

Tabela 3.8 - 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 |

Tabela 3.9 - Vetor BOW após a transformação para minúsculas

Como podemos verificar, existe uma diminuição do vocabulário que, consequentemente, diminui significativamente o tamanho dos vetores, facilitando a capacidade de aprendizagem e convergência dos modelos supervisionados. Na figura 3.7, podemos observar claramente o efeito que o processo de transformação para minúsculas tem no tamanho do vocabulário.

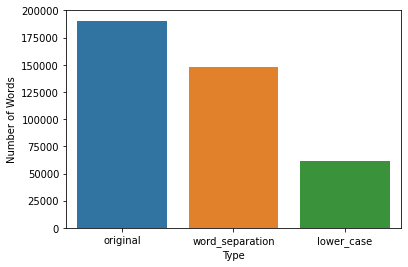


Figura 3.7 – Tamanho do vocabulário após transformação para minúsculas

### *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 [60]. Na tabela 3.10 podemos observar um exemplo do processo de POS Tagging, seguido de uma remoção de todas as palavras das categorias mencionadas anteriormente:

|  |  |
| --- | --- |
| **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”*] |

Tabela 3.10 – Processo de POS Tagging & remoção de palavras

A figura 3.8 mostra-nos a redução do tamanho do vocabulário após a execução deste processo:

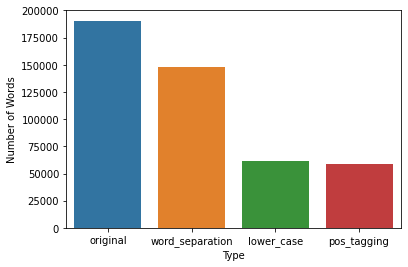


Figura 3.8 - Tamanho do vocabulário após POS Tagging & remoção de palavras

Na realidade, este não é o processo mais comum para remover palavras que não têm conteúdo emocional, sendo que o mais comum é o processo de remoção de *stop-words*. Em primeiro lugar, o processo de remoção de *stop-words* 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 *stop-words*, 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 *stop-words*, 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.

### *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 3.11 mostra alguns exemplos de lemmatization aplicado a algumas palavras:

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

Tabela 3.11 - Exemplos de lemmatization

Eis o resultado da aplicação de lemmatization a uma opinião:

|  |  |
| --- | --- |
| **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”*] |

Tabela 3.12 - Exemplo de aplicação de lemmatization a uma opinião

E, finalmente, a figura 3.9, mostra-nos o tamanho do vocabulário após a aplicação de *lemmatization* ao conjunto dados:

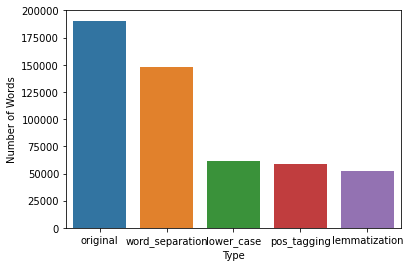


Figura 3.9 - Tamanho do vocabulário após aplicação de *lemmatization*

## 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 3.10, 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.

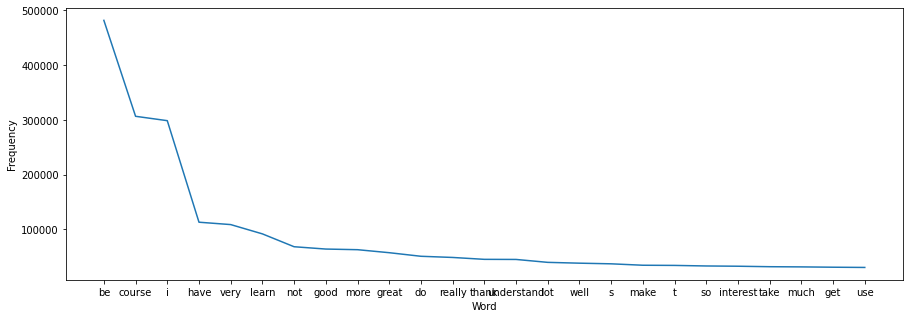


Figura 3.10 – Número de ocorrências das 25 palavras mais comuns após pré-processamento

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. Voltando à figura 3.9 e à tabela 3.3, 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 *dataset* de treino diferente, logo o pré-processamento pode, e deve, ser alterado tendo em conta os resultados dos modelos.

Capítulo 4

# Classificação de Sentimentos

## O Processo de Classificação

De acordo com o que já foi analisado no capítulo de revisão bibliográfica, 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, foi estruturado o seguinte processo de classificação inicial:

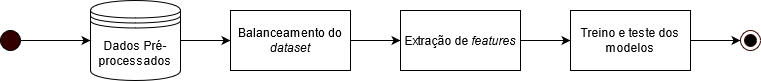


Figura 4.1 - Processo de classificação inicial

Como já foi mencionado anteriormente, o *dataset* 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 *dataset* 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 *features* 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 nos capítulos seguintes, este processo revelou-se insuficiente. A necessidade de classificar as opiniões num valor entre 1 e 5 aumenta significativamente a complexidade do problema, o que resulta 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 uma *feature* extra, algo que será explicado extensivamente no capítulo 4.4. Assim, o processo apresentado na figura 4.1 foi adaptado para o seguinte:

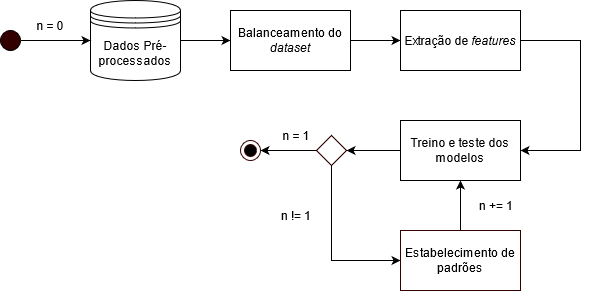


Figura 4.2 – Processo de classificação final

A grande diferença entre o processo inicial e o final é, de facto, a utilização de padrões sintáticos para melhoria dos modelos. Como podemos observar, inicialmente é feito o treino e o teste dos modelos utilizando apenas o processo da figura 4.1, 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.

Nos subcapítulos seguintes este processo será então dissecado, apresentando no geral as técnicas e os modelos desenvolvidos bem como a análise dos resultados dos variados modelos.

## 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 features 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 4.1 apresenta os diferentes modelos desenvolvidos numa fase inicial, bem como as técnicas utilizadas nestes:

|  |  |  |
| --- | --- | --- |
| **Técnicas de balanceamento do *dataset*** | **Modelos de extração de *features*** | **Classificadores** |
| *Undersampling* | Word2Vec | Naive Bayes |
| Random Forest |
| XG Boost |
| Support Vector Machines |
| Bag-of-Words | Naive Bayes |
| Random Forest |
| XG Boost |
| Support Vector Machines |

Tabela 4.1 - Técnicas e modelos desenvolvidos

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, onde será feito um estabelecimento de padrões sintáticos de forma a tentar aumentar a eficácia dos classificadores. Esta escolha será baseada na *accuracy* e 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.

## Um Primeiro Modelo de Classificação

Como já foi explicitado na figura 4.1, o primeiro passo do modelo de classificação consiste em balancear o *dataset*. 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, onde 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 *overfitting*. 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 *dataset* 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 |

Tabela 4.2 - Nº de opiniões por classe

Agora que temos um *dataset* balanceado, é necessário traduzir este para um formato que seja facilmente compreendido por modelos de aprendizagem automática. É possível alcançar isto através da utilização de técnicas de extração de *features* como é o caso das técnicas TF-IDF e Word2Vec.

## Utilização de Padrões Sintáticos

Após analisar os resultados dos primeiros modelos, que serão apresentados no subcapítulo 4.6, foi possível depreender que a eficácia destes não era, de todo, a pretendida. Sendo assim, era necessário fazer alguns ajustes, ajustes estes que passam por disponibilizar mais informação relevante aos modelos, de forma a tentar melhorar a eficácia destes.

Em 1992, *Hearst M.* [] 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 padrões sintáticos nesta, talvez também seja 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 *dataset* foi partido em dois *datasets*, um de treino e outro de teste. Apenas o *dataset* de treino foi utilizado na identificação de padrões sintáticos para que o modelo não fique enviesado quando receber como input o *dataset* de teste.
3. No *dataset* de treino é então utilizada a coluna do *dataframe* que contém a informação sobre a classe gramatical das palavras pertencentes às opiniões. São então retiradas sequências de cinco, seis e sete classes gramaticais consecutivas, sendo estas separadas tendo em conta a classificação da opinião respetiva.
4. Logo de seguida são apenas guardados alguns padrões para cada classe. Esta seleção é feita através do seguinte critério: se para a classe X, a matriz de confusão demonstrar uma grande quantidade de dados incorretamente anotados como classe Y ou Z, serão retirados dos padrões da classe x todos os padrões que também ocorram na classe y ou na classe Z. Tal como será possível observar no subcapítulo 4.6, 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 são criadas cinco colunas extra aos vetores criados pelo modelo TF-IDF. Cada coluna terá 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 melhor facilidade em depreender estes valores.
6. Finalmente são 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:

|  |  |  |  |
| --- | --- | --- | --- |
| **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*” |

Tabela 4.3 – Padrões identificados e as suas respetivas classes, nº ocorrências e exemplos de frases. (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 tabela 4.2, 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, podemos inferir que, em alguns casos, estes padrões são de facto uma ferramenta coerente no que diz respeito ao apoio que dão aos modelos de classificação.

## Índices para a Categorização de Sentimentos

Aqui.

## Análise de Resultados

Aqui.

Capítulo 5

# Conclusões e Trabalho Futuro

## Conclusões

Aqui.

## Trabalho Futuro

Aqui.

# Referências

1. Stiglitz, Joseph E, Amartya Sen, and Jean-Paul Fitoussi. 2009. “Report by the Commission on the Measurement of Economic Performance and Social Progress.”
2. Fleurbaey, M. 2012. “Beyond GDP: The Quest for a Measure of Social Welfare. Part I.” Voprosy Ekonomiki 2012 (2): 67–93. https://doi.org/10.32609/0042-8736-2012-2-67-93.
3. Cummins, Robert A., Richard Eckersley, Julie Pallant, Jackie Van Vugt, and Roseanne Misajon. 2003. “Developing a National Index of Subjective Wellbeing: The Australian Unity Wellbeing Index.” Social Indicators Research 64 (2): 159–90. https://doi.org/10.1023/A:1024704320683.
4. Iacus, Stefano Maria, Giuseppe Porro, Silvia Salini, and Elena Siletti. 2015. “Social Networks, Happiness and Health: From Sentiment Analysis to a Multidimensional Indicator of Subjective Well-Being.” http://arxiv.org/abs/1512.01569.
5. Mathers, Colin D., and Dejan Loncar. 2006. “Projections of Global Mortality and Burden of Disease from 2002 to 2030.” PLoS Medicine 3 (11): 2011–30. https://doi.org/10.1371/journal.pmed.0030442.
6. Twenge, Jean M. 2000. “The Age of Anxiety? Birth Cohort Change in Anxiety and Neuroticism, 1952-1993.” Journal of Personality and Social Psychology 79 (6): 1007–21. https://doi.org/10.1037/0022-3514.79.6.1007.
7. Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. 2014. “Sentiment Analysis Algorithms and Applications: A Survey.” Ain Shams Engineering Journal 5 (4): 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011.
8. Feldman, Ronen. 2013. “Techniques and Applications for Sentiment Analysis.” Communications of the ACM. https://doi.org/10.1145/2436256.2436274.
9. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. “Thumbs Up?” In . https://doi.org/10.3115/1118693.1118704.
10. Turney, Peter D. 2001. “Thumbs up or Thumbs Down?” In . https://doi.org/10.3115/1073083.1073153.
11. Yu, Hong, and Vasileios Hatzivassiloglou. 2003. “Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences.” Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing.
12. Hatzivassiloglou, Vasileios, Judith L Klavans, Melissa L Holcombe, Regina Barzilay, Min-Yen Kan, and Kathleen R. McKeown. 2001. “SIMFINDER : A Flexible Clustering Tool for Summarization.” In Proceedings of the NAACL Workshop on Automatic Summarization.
13. Kim, Soo Min, and Eduard Hovy. 2007. “Crystal: Analyzing Predictive Opinions on the Web.” In EMNLP-CoNLL 2007 - Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.
14. Pontiki, Maria, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, et al. 2016. “SemEval-2016 Task 5: Aspect Based Sentiment Analysis.” In SemEval 2016 - 10th International Workshop on Semantic Evaluation, Proceedings. https://doi.org/10.18653/v1/s16-1002.
15. Hai Z, Chang K, Kim J (2011) Implicit feature identification via co-occurrence association rule mining. In: in: Gelbukh a.F. (eds) computational linguistics and intelligent text processing. CICLing 2011. Lecture notes in computer science, vol 6608. Springer, Berlin, Heidelberg. pp 393–404 (!!!!!!)
16. Ravi, Kumar, and Vadlamani Ravi. 2015. “A Survey on Opinion Mining and Sentiment Analysis: Tasks, Approaches and Applications.” Knowledge-Based Systems. https://doi.org/10.1016/j.knosys.2015.06.015.
17. Tsur, Oren, Dmitry Davidov, and Ari Rappoport. 2010. “ICWSM - A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews.” In ICWSM 2010 - Proceedings of the 4th International AAAI Conference on Weblogs and Social Media.
18. Demirtas, Erkin. 2013. “Cross-Lingual Sentiment Analysis with Machine Translation.” Utility of Training Corpora and Sentiment Lexica, Master Thesis, University of Technology.
19. Boiy, Erik, and Marie Francine Moens. 2009. “A Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts.” Information Retrieval. https://doi.org/10.1007/s10791-008-9070-z.
20. Tan, Songbo, Xueqi Cheng, Yuefen Wang, and Hongbo Xu. 2009. “Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis.” In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-642-00958-7\_33.
21. Weichselbraun, Albert, Stefan Gindl, and Arno Scharl. 2013. “Extracting and Grounding Contextualized Sentiment Lexicons.” IEEE Intelligent Systems. https://doi.org/10.1109/MIS.2013.41.
22. Vinodhini, G, and RM Chandrasekaran. 2012. “Sentiment Analysis and Opinion Mining : A Survey International Journal of Advanced Research in Sentiment Analysis and Opinion Mining : A Survey.” International Journal of Advanced Research in Computer Science and Software Engineering.
23. Kang, Hanhoon, Seong Joon Yoo, and Dongil Han. 2012. “Senti-Lexicon and Improved Naïve Bayes Algorithms for Sentiment Analysis of Restaurant Reviews.” Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2011.11.107
24. Basari, Abd Samad Hasan, Burairah Hussin, I. Gede Pramudya Ananta, and Junta Zeniarja. 2013. “Opinion Mining of Movie Review Using Hybrid Method of Support Vector Machine and Particle Swarm Optimization.” In Procedia Engineering. https://doi.org/10.1016/j.proeng.2013.02.059.
25. O’Connor, Brendan, Michel Krieger, and David Ahn. 2010. “TweetMotif: Exploratory Search and Topic Summarization for Twitter.” In ICWSM 2010 - Proceedings of the 4th International AAAI Conference on Weblogs and Social Media.
26. Kannan, S., Vairaprakash Gurusamy, S. Vijayarani, J. Ilamathi, Ms. Nithya, S. Kannan, and Vairaprakash Gurusamy. 2015. “Preprocessing Techniques for Text Mining.” International Journal of Computer Science & Communication Networks.
27. Jivani, Ms. Anjali Ganesh. 2007. “A Comparative Study of Stemming Algorithms.” October.
28. Maks, Isa, and Piek Vossen. 2012. “A Lexicon Model for Deep Sentiment Analysis and Opinion Mining Applications.” In Decision Support Systems. https://doi.org/10.1016/j.dss.2012.05.025.
29. Pang, Bo, and Lillian Lee. 2004. “A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts.” https://doi.org/10.3115/1218955.1218990.
30. Bravo-Marquez, Felipe, Marcelo Mendoza, and Barbara Poblete. 2014. “Meta-Level Sentiment Models for Big Social Data Analysis.” Knowledge-Based Systems. https://doi.org/10.1016/j.knosys.2014.05.016.
31. Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. “SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.” In Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010.
32. Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings, 1–12.
33. Chen, Min, Shiwen Mao, and Yunhao Liu. 2014. “Big Data: A Survey.” In Mobile Networks and Applications. https://doi.org/10.1007/s11036-013-0489-0.
34. Montoyo, Andrés, Patricio Martínez-Barco, and Alexandra Balahur. 2012. “Subjectivity and Sentiment Analysis: An Overview of the Current State of the Area and Envisaged Developments.” In Decision Support Systems. https://doi.org/10.1016/j.dss.2012.05.022.
35. Le, Quoc, and Tomas Mikolov. 2014. “Distributed Representations of Sentences and Documents.” 31st International Conference on Machine Learning, ICML 2014 4: 2931–39.
36. Mccallum, Andrew, and Kamal Nigam. 1998. “A Comparison of Event Models for Naive Bayes Text Classification.” AAAI-98 Workshop on Learning for Text Categorization 752: 41–48.
37. Joachims, Thorsten. 1998. “Text Categorization with Support Vector Machines: Learning with Many Relevant Features.” In , 137–42. https://doi.org/10.1007/BFb0026683.
38. Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” Advances in Neural Information Processing Systems 2017-December (Nips): 5999–6009.