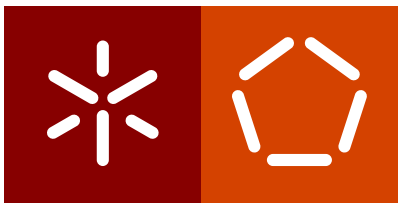
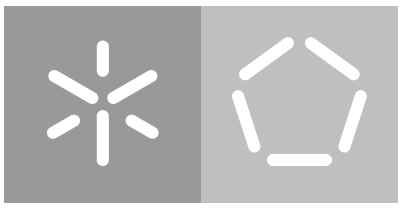


Universidade do Minho
Escola de Engenharia
Departamento de Informática

João Miguel Ferreira Lopes

Análise de Sentimentos em Conteúdos Textuais

Dezembro 2018



Universidade do Minho
Escola de Engenharia
Departamento de Informática

João Miguel Ferreira Lopes

Análise de Sentimentos em Conteúdos Textuais

Tese de Mestrado
Mestrado em Engenharia Informática

Trabalho efetuado sob a orientação de
Professor Doutor Orlando Manuel de Oliveira Belo
Investigador Auxiliar Pedro Gabriel Dias Ferreira

Dezembro 2018

Agradecimentos

Não poderia deixar de iniciar esta pequena parte que me é reservada para prestar tributo às pessoas que fizeram parte do caminho até aqui tomado sem começar por agradecer de forma especial à minha família. Aos meus pais que tudo fizeram e fazem para me proporcionarem a melhor vida e educação possível e sobretudo apoio e amor incondicional, pelos sacrifícios a que se sujeitam e que jamais serei capaz de retribuir, não existe uma palavra suficientemente forte para exprimir o quanto agradeço todo o que fizeram e fazem por mim. Ao meu irmão que sempre me ensinou desde pequeno, que sempre admirei e suscitou em mim o interesse pelo estudo, aprendizagem e que foi uma inspiração para as decisões que tomei e que me trouxeram até aqui, pelas brincadeiras e zangas de irmãos, muito obrigado.

Não podia deixar também de agradecer a uma pessoa muito especial na minha vida, a minha namorada com quem partilho 11 anos e que sempre esteve do meu lado. A dedicação, apoio e amor que recebo foram fundamentais para conseguir alcançar esta fase. As muitas noites tardias e muitos fins-de-semana a estudar e a fazer trabalhos, contigo ao meu lado pareciam apenas minutos, obrigado por sempre me apoiares. Aproveito ainda para agradecer à tua família que sempre me receberam como se tal fosse, fazendo da tua casa como minha.

Também não podia deixar de agradecer ao meu orientador, Professor Doutor Orlando de Oliveira Belo e ao meu coorientador Pedro Gabriel Dias Ferreira pela orientação nos trabalhos efetuados nesta tese de mestrado. Ao Professor Orlando Belo, durante os anos até este momento, muitos professores foram responsáveis pela minha educação e alguns ficarão sempre presentes ao longo dos anos como é o seu caso, pela forma de ensino exemplar, dedicação, disponibilidade e preocupação pelo desempenho e aprendizagem dos seus alunos, obrigado.

Por último, resta-me agradecer a todas as restantes pessoas que me apoiaram e ajudaram a atingir esta meta, levo comigo muitos amigos sem os quais o percurso poderia ser muito diferente. Também um agradecimento a todas as pessoas que constituem o departamento de informática, pela dedicação e interesse mostrado aos alunos, pela aprendizagem com problemas desafiantes que agora são relembrados como os mais marcantes e que permitem desenvolver competências novas para além do expectável e pela disponibilidade geral que mostram na resolução de qualquer problema.

Abstract

Text Sentiment Analysis

The huge amount of data that is generated on a daily basis by companies and individuals has raised the interest of entities that saw the oportunities in exploiting that information. Soon, the development of data analysis solutions started to emerge rapidly and dynamically. In most cases, these forms of exploiting data are used by profiling systems, as a way of feeding them relevant data, in order to establish some behavioral pattern.

Sentiment analysis in texts is a field of data analysis which has raised much interest in recent years, having been gradually applied over a wide range of problems in order to determine, for example, how a given product is being accepted by people. However, although there are already several models developed for this type of text analysis, their accuracy is still much questioned, in some way due to the difficulties that exist in the accomplishment of this type of analysis in which, in a certain way, it is necessary that written language can be understood, in a natural way, by a given set of algorithms.

In this dissertation this aspect of data analysis will be explored. It is created a set of transformations that are applied to the data, which is in textual format, representative of the pre-processing to be applied in order to transform the data into an appropriate format to be processed by the models. Throughout the preprocessing construction it is also demonstrated the importance of this phase, for any data analysis problem, that without it, it is not possible to understand the analysis problem and the results obtained are not the best possible. Once the data is pre-processed, it is formed a set of models that use techniques of text analysis with the aim of recognizing feelings in it. These models can be summarized to three main ones: supervised model of machine learning, model based on dictionaries and a hybrid model. In any of the models it is sought to extract the maximum possible amount of information, besides the recognition of feelings and its polarity, as recognition of the aspects to which the feelings refer, among others. Of the three models developed, the hybrid model was the one that obtained the best results, with a percentage of incorrect classifications approximately equal to 6% of the total of the test data.

Resumo

Análise de Sentimentos em Conteúdos Textuais

A grande quantidade de dados que é gerada diariamente em empresas ou por pessoas em termos individuais despertou a atenção de algumas entidades que viram o grande interesse e potencial da exploração dessa informação. O desenvolvimento de soluções orientadas para esse tipo de exploração começou, assim, a ser incentivado de forma muito dinâmica. Na maioria dos casos, essa exploração tem como objetivo alimentar sistemas de *profiling*, que posteriormente tentam estabelecer algum tipo de padrão comportamental através da utilização de uma ou mais técnicas de análise de dados.

A análise de sentimentos presentes em textos é uma das áreas de análise de dados que também tem despertado muito interesse nos últimos anos, tendo sido gradualmente aplicada sobre uma gama de problemas muito diversificada para determinar, por exemplo, como é que um dado produto está a ser aceite pelas pessoas. Contudo, embora existam já vários modelos desenvolvidos para este tipo de análise, a sua precisão ainda é muito questionada, em parte devido às dificuldades que existem na realização deste tipo de análise, na qual, de certa forma, é necessário que a linguagem escrita seja compreendida de forma natural por um dado conjunto de algoritmos.

Neste trabalho de dissertação explorámos esta vertente de análise de dados, com particular ênfase na análise de sentimentos em conteúdos textuais. Foi aplicado um conjunto de transformações responsáveis pelo pré-processamento e transformação dos dados para um formato apropriado para serem utilizados pelos modelos. Ao longo da construção do pré-processamento foi, ainda, demonstrada a importância desta fase, para qualquer problema de análise de dados, que sem ela não é possível compreender o problema de análise o que frequentemente leva a que os resultados obtidos não sejam os melhores possíveis. Após o pré-processamento dos dados, foram desenvolvidos três modelos de análise de sentimentos em textos: *modelo supervisionado de aprendizagem automática*, *modelo baseado em dicionários de sentimentos* e *modelo híbrido*. Qualquer um dos modelos faz uso de técnicas de análise de textos de modo a serem reconhecidos sentimentos e respetivas polaridades, aspetos a que os sentimentos se referem,

entre outros. Dos três modelos desenvolvidos, o modelo híbrido foi o que obteve melhores resultados, com uma percentagem de classificações incorretas aproximadamente igual a 6% do total dos dados de teste.

Conteúdo

Lista de Figuras

Lista de Tabelas

1 Introdução

1.1 Enquadramento de Geral

A extração de padrões e de conhecimento a partir de dados ocorre desde à muitos anos (1700s). Desde essa altura, à medida que nos aproximamos da atualidade, o poder computacional aumentou exponencialmente bem como a quantidade de dados gerados por cada entidade. O aumento crescente da capacidade tecnológica e da digitalização de todo o tipo de informação fez com que diariamente sejam gerados grandes conjuntos de dados, de natureza muito diversa. Ao mesmo tempo que os dados iam aumentando, aumentou também a quantidade de ruído presente neles. Sistematicamente a formação de novos conjuntos de dados e a sua recolha para um único local, como sucede com os *data warehouses (DWs)*, está frequentemente sujeita a erros, tanto humanos como computacionais, que podem comprometer a qualidade dos dados em futuras utilizações. Tem sido cada vez mais importante para empresas, setores do estado e entidades individuais conseguir diferenciar os dados por eles gerados ou recolhidos, de forma a se poder obter apenas aqueles que realmente são importantes e que permitam inferir alguma informação útil acerca da área de análise em que se enquadram. Ao longo do tempo, e à medida que a quantidade de dados ia aumentando, foram desenvolvidas várias soluções computacionais, cada vez mais eficazes no tratamento de dados, para a obtenção de conhecimento útil acerca das atividades desenvolvidas.

As empresas e outro tipo de entidades necessitam cada vez mais de implementar sistemas de *profiling* com o intuito de extrair informação relevante a partir dos dados produzidos nas atividades por elas desenvolvidas. Usualmente, estes sistemas de *profiling* utilizam métodos computacionais de tratamento e de obtenção de informação a partir de conjuntos de dados, por forma a representar de algum modo uma pessoa ou um conjunto de pessoas. O *Profiling* não se limita, contudo, à representação de conhecimento sobre pessoas, podendo ser aplicado em áreas como, por exemplo, a deteção de fraudes, a segurança, a atribuição de créditos, entre muitos outras, graças à qualidade de tratamento e associação de informação representativa de cada área. A utilização de *profiling* tem ainda aumentado de forma acentuada nos últimos anos à medida que as tecnologias móveis são cada vez mais utilizadas, por um número de utilizadores cada vez maior, uma vez que este tipo de tecnologias são, usualmente, de natureza pessoal

e conseguem gerar grandes quantidades de dados relativos às preferências e hábitos de cada utilizador.

Na construção de um sistema de *profiling* sobre uma dada área de análise é necessário assegurar que toda a informação e recursos necessários para o processo se encontram disponíveis. A construção destes sistemas segue um procedimento geral que garante que todos os recursos necessários se encontram disponíveis, como por exemplo:

- **A determinação do problema** – definir inicialmente qual o problema a tratar e os objetivos a atingir.
- **A recolha de dados** – assegurar o acesso às fontes de dados necessárias para que seja possível obter o conjunto de dados relevantes para análise do problema.
- **A preparação dos dados** – o processo de preparação e de limpeza dos dados recolhidos. Esta etapa á crucial uma vez que garante que os dados se encontram no formato necessário e apenas aqueles que f orem realmente úteis são utilizados para análise sem que ocorra introdução de ruído.
- **A mineração de dados (*data mining*)** – modelação do conjunto de dados com a aplicação de modelos e técnicas de análise e de extração de conhecimento utilizando esses dados como base de trabalho. Os modelos aplicados vão de acordo com a natureza dos dados e dos objetivos a atingir.
- **A validação e aplicação** – os resultados obtidos pelo processo de *data mining* são avaliados em relação à sua qualidade e relevância. Após a validação dos modelos desenvolvidos, estes são aplicados sobre o domínio para o qual foram desenvolvidos de modo a testar e validar os modelos em situações reais e otimizar os mesmos caso necessário.

Embora todas as etapas da construção de um *profile* sejam essenciais para o resultado final e da sua aplicação, em grupos externos aos de teste, a etapa de *data mining* possui uma relevância acrescentada em comparação com as restantes.

O *data mining* é um processo que faz uso de técnicas de obtenção de conhecimento que permitem encontrar padrões em conjuntos de dados, permitindo que seja extraída e transformada algum tipo de informação que possa ser utilizada no futuro. O *data mining* envolve nos seus processos a utilização de grandes volumes de dados, bem como várias aplicações especialmente orientadas para suporte à decisão. O *data mining* pode ser visto como um processo independente que possui várias aplicações externas ao processo de *profiling*, sendo amplamente usado em problemas de análise de dados.

De facto, o *data mining* é uma das muitas técnicas utilizadas em análise de dados – a análise exploratória, confirmatória e preditiva de dados e ainda a análise textual são algumas outras técnicas de análise de dados.

1.2 Análise de Sentimentos em Textos

Hoje em dia existem muitas formas de apresentar e armazenar informação, sendo uma delas, provavelmente a mais antiga, informação em formato textual. De facto, a representação e a difusão de conhecimento em formato escrito mantém-se, ainda hoje, como o formato mais utilizado. De entre o vasto número de diferentes linguagens utilizadas a nível mundial e da evolução que cada uma delas sofreu desde a sua origem até à atualidade, é fácil compreender a grandeza da informação presente em textos ou que possa estar contida em livros com informação dos nossos antepassados, até simples frases de opiniões, em formato digital, espalhadas pela Internet. Assim, a possibilidade de extração de conhecimento de forma automática a partir de textos tem grande importância, especialmente à medida que a quantidade de informação aumenta exponencialmente com o auxílio de novas tecnologias. A filtragem de certos tipos de conteúdos textuais específicos para análise e extração automática de conhecimento neles contido, permite que esse conhecimento seja aplicado por empresas, ou outro tipo de entidades, nas atividades operacionais mundanas ou em processos de tomada de decisão em conformidade com a informação recolhida.

A análise de sentimentos a partir de textos é uma das muitas atividades de análise que se pretende ver realizada de forma automática e eficaz. Este tipo de análise orientada para textos visa determinar o "estado de espírito" dos seus autores que, por muitas vezes, de alguma forma, é refletido na forma de escrita e no conteúdo dos textos – de relevar aqui que a noção de texto pode ser sustentada por argumentos provenientes desde grandes parágrafos até pequenas frases. A utilização de técnicas de aprendizagem automática, em conjunção com a aplicação de informação estatística e técnicas de processamento de linguagem natural, são algumas das formas utilizadas para desenvolver uma solução de análise de sentimentos, potenciando a sua automatização, e ser aplicada ao longo do tempo sobre um grande conjunto de alvos, de grande diversidade e natureza distinta, que poderão ser conteúdos de páginas Web, notícias e opiniões online, entre muitos outros. Todos estes tipos de conteúdos têm aberto um grande espaço de análise, aumentando de forma quase exponencial a necessidade e a procura de soluções para a realização deste tipo de análise por parte de empresas, estabelecimentos, organismos estatais, entre outros. Com base nos resultados desses processos, todas essas entidades esperam poder agir e tomar decisões de acordo com o sentimento que as pessoas vão expressando, sobre este ou aquele tema, sobre este ou aquele produto. Porém, a aplicação de processos de análise

de sentimentos não se limita simplesmente a atuar sobre conteúdos online, sendo também aplicado, cada vez mais, sobre sistemas empresariais em processos de obtenção de informação pertinente, para ajudar no desenvolvimento dos seus modelos de negócio de forma sustentada.

O desenvolvimento de técnicas de análise de sentimentos não é um processo simples uma vez que implica o desenvolvimento e aplicação de algoritmos bastante sofisticados (e complexos), que tenham a capacidade de compreender de forma natural a linguagem escrita. Essa dificuldade acentua-se um pouco mais quando são considerados vários fatores culturais, bem como a presença de diferentes tons de escrita, que a podem influenciar, tornando por vezes difícil o processo de interpretação da informação mesmo quando efetuado por pessoas. Torna-se, portanto, relevante a exploração das várias técnicas existentes para que seja possível evoluir na interpretação dos textos e aumentar a capacidade de extração de conhecimento a partir deles.

A análise de sentimentos em textos é uma área recente. Antes do ano 2000 este conceito de análise ainda não se encontrava totalmente presente na comunidade de investigação, havendo pouca investigação ou interesse na área. Antes desta data, os primeiros trabalhos que se podem considerar ser próximos da análise de sentimentos eram baseados no processamento da linguagem natural para interpretação de metáforas, estudo de linguagem natural ou na identificação de emoções e pontos de vista. Foi em 2001 que a generalização do conceito de análise de sentimentos começou a ser adotado bem como a identificação das áreas de interesse da sua aplicação. O motivo pelo qual este ano marcou a expansão da área deve-se em grande parte ao desenvolvimento e crescimento dos conteúdos online que permitiram obter grandes conjuntos de dados para análise, em que as opiniões das pessoas acerca de todo o tipo de temas começaram a proliferar-se. Para além disto, a investigação até ali efetuada possibilitou o crescimento de modelos de aprendizagem automática e de métodos de processamento de linguagem natural. À medida que estes fatores cresciam originavam um crescente despertar de um maior interesse sobre as dificuldades que a área coloca e das possibilidades de aplicação no meio empresarial e individual.

Atualmente, a análise de sentimentos conseguiu alcançar um vasto número de domínios de aplicação, desde a análise de produtos ou de serviços médicos e financeiros [18]. Muitas das principais empresas do mercado tecnológico e de software desenvolveram soluções internas de análise de sentimentos, como foram os casos da Microsoft e da Google. Na literatura, a análise de sentimentos é vastamente aplicada sobre filmes [5] e produtos. Todavia, outras aplicações são também estudadas, por exemplo a previsão dos resultados eleitorais e previsão do mercado de ações [8].

1.3 Motivação e Objetivos

A proliferação crescente que nos últimos anos se tem verificado na digitalização de conteúdos, aumentou de forma muito acentuada a necessidade de saber como explorar a informação neles contida, com vista à aquisição de conhecimento que possa contribuir para a obtenção de vantagens em qualquer ambiente de decisão e de negócios. A utilização de técnicas de *profiling* foi uma das formas encontradas para ajudar as empresas, e outras entidades, na identificação e caracterização de aspetos críticos relacionados com as suas atividades. Em muitos casos, esses aspetos estão associados com a perceção e opinião que uma pessoa, ou conjunto de pessoas, possui sobre um determinado tema, ou seja, o sentimento que essas pessoas revelam face ao alvo de análise, quer este seja um serviço ou um produto. Tal sentimento pode ser representado em vários formatos sendo um deles o textual. A análise de sentimentos em textos tem sido muito útil em áreas como o *profiling*, ajudando na caracterização de perfis complementando-os com novos elementos provenientes de reações, de opiniões ou de sentimentos expressos em textos.

Uma pessoa tem a capacidade de ler um conjunto de informação textual e identificar os sentimentos expressos pelo autor do texto acerca dos vários tópicos que nele são abordados. Contudo, não raras vezes, mesmo para uma pessoa, este processo pode ser complicado, devido à grande subjetividade relacionada com os sentimentos e a forma como estes são expressos. Por exemplo, para o autor do texto, o formato de escrita é representativo de um dado grau de intensidade do sentimento, mas para o leitor, que possui um outro conhecimento sobre as diversas formas de escrita, esse grau pode ser muito menor ou maior, podendo originar algum tipo de incerteza acerca do sentimento presente no texto. Este problema agrava-se quando o processo de análise de sentimentos é realizado por um computador, o qual, naturalmente, não possui conhecimento natural acerca da linguagem. É, então, necessário uma estruturação do documento de análise de modo a ser representado numa dada estrutura, que possa ser interpretada pelos modelos computacionais que fazem uso de algoritmos que, de alguma forma, permitam associar um dado contexto e um sentimento a cada palavra ou conjunto de palavras contidas no documento de análise.

Vários modelos foram desenvolvidos para análise de sentimentos, contudo, todos eles, uns mais do que outros, sofrem de problemas como:

- **A deteção de sarcasmo** – é extremamente difícil para um computador detetar sarcasmo, por forma a identificar o sentimento oposto.
- **A resolução de pronomes** – opiniões e sentimentos sobre algo que seja representado por um pronome torna difícil identificar o alvo desse pronome.

- **A atribuição da força dos sentimentos** – dependendo da forma de escrita, uma opinião pode ser forte ou fraca mesmo quando tem a mesma polaridade.

- **As palavras de contraste** – informação com conjunções especialmente se derem oportunidade a opiniões distintas. Ex.: "Adorei o filme mas não gostei das cadeiras".

Estas são apenas algumas das limitações e problemas atuais que podemos encontrar nos métodos de análise de sentimentos – noutras secções abordaremos outros.

Neste trabalho de dissertação procurou-se analisar alguns dos modelos e técnicas existentes no domínio da análise de sentimentos, tendo como motivação o estabelecimento de um sistema que permitisse aplicar um modelo desenvolvido sobre um domínio de aplicação concreto, que envolva a definição e caracterização de perfis de comportamento. Mais concretamente, procurou-se:

- Explorar técnicas de análise e qualidade de dados em formato textual.

- Explorar e construir padrões de pré-processamento de dados textuais com o objetivo de os transformar de modo adequado para a identificação de sentimentos e outra informação relevante neles presentes.

- Construir modelos de aprendizagem automática para a análise de sentimentos em textos com recurso a várias metodologias frequentemente utilizadas neste tipo de problemas.

- Construir modelos de análise de sentimentos em textos baseados em dicionários de sentimentos.

- Analisar os resultados obtidos e comparar a qualidade dos modelos desenvolvidos.

- Resolver de problemas e dificuldades comuns à análise de sentimentos em textos.

1.4 Estrutura da Dissertação

Para além deste capítulo, esta dissertação inclui mais 7 capítulos, a saber:

- **Capítulo 2 – Análise de Sentimentos.**
  Neste capítulo é apresentada informação sobre a análise sistemática sobre o estado da arte em análise de sentimentos em conteúdos textuais e as várias técnicas de processamento de textos, bem como os problemas e as soluções encontradas para os ultrapassar.

- **Capítulo 3 – Pré-Processamento dos Dados.**
  Aqui são apresentados vários métodos de análise do conjunto de dados para a sua compreensão, em conjunto com o desenvolvimento da aplicação de técnicas de pré-processamento responsáveis por transformar os dados em formatos apropriados para utilização nos vários modelos desenvolvidos.

- **Capítulo 4 – Construção de Recursos.**
  Neste capítulo é apresentado detalhadamente os dois recursos principais (de entre outros) que foram desenvolvidos para serem utilizados pelos modelos de aprendizagem automática bem como pelo modelo baseado em dicionários. Estes recursos são, nomeadamente, o *dicionário de sentimentos* e o *POS Tagger*. O primeiro é um léxico de correspondência entre palavra - polaridade - POS tag. O segundo é um modelo de marcação de partes do discurso capaz de marcar automaticamente uma palavra com a sua respetiva POS tag tendo em conta o contexto envolvente da mesma.

- **Capítulo 5 – Modelos Supervisionados.**
  Tendo os dados pré-processados, podemos construir a partir de textos modelos de deteção de sentimentos e de previsão da respetiva polaridade. Aqui apresenta-se o processo de construção de um conjunto de modelos supervisionados de aprendizagem automática que fazem uso de técnicas e transformações frequentemente utilizadas em processamento de texto e de linguagem natural. O objetivo é comparar os resultados de cada modelo de modo a sustentar quais as técnicas mais apropriadas para utilização no conjunto de dados utilizado.

- **Capítulo 6 – Um Modelo Baseado em Dicionários.**
  Aqui apresenta-se um modelo que utiliza o dicionário de sentimentos anteriormente construído para obter a polaridade de cada palavra de sentimento reconhecida no texto. O processo na sua forma mais simples pode-se resumir à procura direta no dicionário de uma palavra. Contudo, o modelo não se resume apenas a este processo, tendo sido desenvolvidas várias técnicas para assegurar a correta polaridade, bem como a obtenção de mais informação como o contexto do texto de análise, os aspetos positivos e negativos nele descritos, e ainda o quanto positivo ou negativo é um sentimento nele expresso.

- **Capítulo 7 – Um Modelo Híbrido.**
  O melhor modelo de aprendizagem automática conseguido no capítulo 5 e o modelo baseado em dicionários do capítulo anterior possuem, cada um, as suas vantagens e desvantagens. O modelo desenvolvido neste capítulo visou conciliar as características dos dois modelos anteriores em apenas um, tirando assim partido dos pontos fortes de cada modelo, com o objetivo de, potencialmente, aumentar a qualidade das previsões

em comparação com os modelos anteriores.

- **Capítulo 8 – Conclusões e Trabalho Futuro.**
  Neste capítulo é apresentado um resumo dos modelos desenvolvidos e dos resultados obtidos e um ponto crítico dos mesmos e ainda possíveis pontos futuros de orientação do trabalho para melhorar os resultados.

2 Análise de Sentimentos

2.1 Introdução

A capacidade de expressar sentimentos de forma escrita é um atributo inerente apenas ao ser humano, desenvolvido ao longo de muitos anos de evolução, cada vez mais com maior complexidade na capacidade de comunicação escrita. Ao longo desta evolução são cada vez mais evidentes as formas como os sentimentos são expressos textualmente, desde simples frases de opinião, cujo sentimento é claramente identificado: *"eu gosto disto"*, até frases complexas que fazem uso de opiniões sarcásticas para expressar os sentimentos acerca da entidade de análise. A definição dos termos "sentimentos" e "análise de sentimentos" são de grande importância. Sentimentos são reconhecidos como emoções, julgamentos, opiniões ou ideias induzidas por emoções ou suscetibilidades. Nos sistemas computacionais que lidam com o processamento de sentimentos, o foco recai essencialmente sobre opiniões presentes em textos. A informação presente em textos pode ter duas formas: factos e opiniões. Os factos apresentam informação objetiva acerca de objetos, entidades, eventos ou características, enquanto que as opiniões são expressas por informação subjetiva, pertencente a uma ou mais entidades, que apresenta informação acerca de sentimentos, opiniões, pontos de vista, expressos por pessoas, em relação a objetos, entidades, eventos ou tópicos [21].

Bing Liu [18] define análise de sentimentos, também denominada por análise de opiniões, como o estudo de sentimentos, opiniões, avaliações, apreciações, atitudes e emoções relativamente a entidades como produtos, serviços, organizações, pessoas, problemas, eventos, tópicos e caraterísticas. Outras nomenclaturas como análise de opiniões, extração de opiniões, análise de subjetividade, análise de emoções, análise de polaridade são aplicadas na literatura como sinónimos do processo de análise de sentimentos. Existem, contudo, autores que diferenciam os termos análise de opiniões e análise de sentimentos [33]. No trabalho referido, a análise de opiniões extrai e analisa as opiniões das pessoas acerca de uma entidade, enquanto que análise de sentimentos identifica o sentimento expresso num texto e depois analisa-o. Assim, o objetivo da análise de sentimentos é identificar opiniões, os sentimentos que estas expressam e, por fim, definir a sua polaridade. Apesar desta distinção por parte de alguns autores, os dois termos em conjunto com os anteriormente referidos acima encontram-se universalmente

sobre o teto do termo geral Análise de Sentimentos [18]. A análise de sentimentos é, portanto, um processo computacional de extração de informação subjetiva, presente em textos, referente a uma ou mais entidades, que tem como finalidade a identificação e a classificação dos sentimentos neles presentes. Dada a natureza textual das fontes de extração e análise dos sentimentos, os processos neles envolvidos estão intrinsecamente relacionados com os processos de *processamento de linguagem natural (NLP)*. A análise de sentimentos é, de facto, um processo computacional pertencente à área de NLP. Esta consegue alcançar praticamente todos os aspetos caraterísticos dos problemas de NLP como negações *(negation handling)*, desambiguação do sentido das palavras *(word sence disambiguation)* e resolução de pronomes *(coreference resolution)*, os quais são problemas considerados difíceis, dado ainda não existir uma solução universal. Neste sentido, as duas áreas desenvolvem-se mutuamente, dado que qualquer avanço computacional que ocorra numa vai poder ser aplicado sobre a outra. Contudo, é importante realçar que o problema de análise de sentimentos é um problema restrito da área de NLP uma vez que não faz uso de todos os seus constituintes, uma vez que não necessita analisar e compreender totalmente a semântica de cada frase mas apenas alguns aspetos, como as entidades e características presentes no texto e respetiva polaridade dos sentimentos a elas associados.

2.2 Definição do Problema

Enquanto que, para as pessoas, a perceção de sentimentos expressos em formato textual é, na maioria das vezes, fácil, graças à nossa natureza sentimental e emocional, para um sistema informático esta perceção está longe de atingir um patamar de perceção próximo. Maioritariamente, o processo de análise de sentimentos maioritariamente identifica opiniões, sobre o formato de conteúdo subjetivo, que expressa explicitamente ou implicitamente sentimentos positivos ou negativos. Este carácter subjetivo é, portanto, uma das principais características da análise de sentimentos. Antes de expor a forma de como o conteúdo subjetivo é identificado em textos é necessário entender o alvo central de toda a análise envolvente no processo: as opiniões.

2.2.1 Definição de Opinião

Liu B. [18] definiu formalmente uma opinião como um quíntuplo:

- **Opinião -** $(e_i, a_{ij}, s_{ijkl}, h_k, t_t)$

Na definição acima apresentada, e_i representa uma entidade, a_{ij} é um aspeto da entidade e_i, s_{ijkl} é o sentimento expresso sobre o aspeto a_{ij} da entidade e_i, h_k é o autor da opinião e t_t é a data em que a opinião foi expressa pelo autor. De uma forma menos formal, podemos dizer que uma opinião é expressa por um autor, que é o detentor da opinião, que expressa sentimentos acerca dos aspetos de uma entidade num dado momento. Além disso, *Liu B.* [18] fez ainda algumas observações importantes referentes à definição apresentada, nomeadamente:

- Os sentimentos expressos podem ser positivos, negativos ou neutros. Na maioria das aplicações interessam apenas os sentimentos positivos e os negativos, mas os sentimentos neutros também podem ser úteis.

- A classificação dos sentimentos como positivos, negativos ou neutros não é a única forma de classificação. Outras podem ser aplicadas como uma escala nominal de 1 a 5 por exemplo.

- Quando uma opinião é expressa sobre a entidade como um todo, ou seja, não são especificados aspetos, é utilizado o termo *GERAL* para denominar o aspeto.

- Os cinco componentes da definição são essenciais. A falta de um deles pode comprometer a utilidade do processo. Por exemplo, a falta da componente temporal impossibilita a análise de opiniões ao longo do tempo, dado que uma opinião expressa anos atrás é menos relevante do que uma opinião atual.

- A definição é aplicável para a maioria do conteúdo de opinião mas não consegue cobrir todas as formas de opiniões nem diferenciar contextos. Maioritariamente, é aplicável ao tipo de opiniões regulares, mas não é aplicável a opiniões comparativas as quais necessitam de uma definição diferente.

A definição de opinião apresentada é aplicável em níveis de análise de sentimentos muito completos, nos quais se vai ao nível da identificação dos aspetos da entidade. Existem, contudo, outros níveis de análise mais relaxados nos quais não é necessário descer a um nível tão específico como os aspetos. Nestes casos, a definição também pode ser relaxada ignorando a componente aspeto. Outras componentes também podem ser ignoradas, consoante os requisitos do sistema com o possível custo de perda de informação potencialmente relevante para análise no futuro.

2.2.2 O Processo Geral de Análise de Sentimentos

A implementação de um sistema que consiga identificar e classificar sentimentos de forma eficaz é de extrema dificuldade. Tal sistema não é constituído por uma rotina única que efetua todo o processo mas é constituído por um conjunto de pequenos processos que analisam partes da informação que queremos analisar. O desempenho em cada processo é, portanto, de grande importância dado que basta um obter uma precisão de análise baixa para comprometer a precisão global de todo o sistema.

Dada a definição de opinião e um documento de análise d, um sistema de análise de sentimentos tem como objetivo determinar todos os quíntuplos de d. A análise de sentimentos consiste, então, de uma forma geral na execução de um conjunto de processos (Figura 2.1).

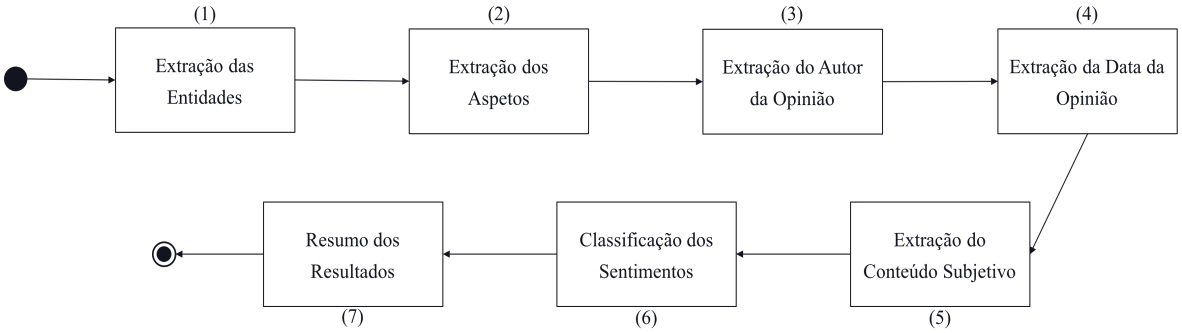


Figura 2.1: Processo geral de análise de sentimentos.

O processo de análise de sentimentos apresentado é o mais completo possível para a maioria dos sistemas de análise de sentimentos. Dependendo do sistema, pode ser necessário acrescentar novos processos ou extrair alguns dos apresentados na Figura 2.1. De facto, se olhar-mos para o processo de forma a reduzi-lo ao mais básico possível, todos os processos apresentados podem ser excluídos, com a exceção dos processos 5, 6 e 7 – a extração do conteúdo subjetivo, a classificação dos sentimentos e o resumo dos resultados, respetivamente. Desta forma, obtém-se um sistema de análise de sentimentos mais básico possível mas que contém os processos de maior dificuldade. A extração de conteúdo subjetivo e classificação dos sentimentos nele presentes são, de facto, os processos que justificam o interesse e o desenvolvimento de soluções de análise de sentimentos. Deste modo, quando se refere que é um sistema básico, está-se a referir ao número de possibilidades de extração de informação que pode ser útil para análise. Todavia, os processos responsáveis por identificar e classificar os sentimentos são de grande complexidade. A definição de cada um dos processos ajuda a compreender melhor o porquê da possibilidade de acréscimo e extração de alguns.

Extração das Entidades, dos Aspetos, do Autor e da Data da Opinião

Dado que em muitos casos um documento contém informação acerca de várias entidades é necessário identificar apenas aquelas que se querem analisar, basicamente são aquelas que representam o tópico de interesse. O processo de extração de entidades é baseado no problema de *named entity recognition (NER)* sobre uma sub-tarefa de extração de informação [18]. Os métodos mais recentes de resolução deste problema são baseados em regras de extração e em modelos estatísticos. As regras de extração baseiam-se num conjunto de regras manualmente implementadas e específicas para um tipo de documento que sejam capazes de extrair a informação necessária, que neste caso são as entidades. Os modelos estatísticos de uma forma geral baseiam-se em *Conditional Random Fields* [31]. Para além destes dois métodos pode ainda ser utilizada aprendizagem automática que utiliza um conjunto de documentos anotados utilizados para treino do modelo de aprendizagem, que é depois utilizado para efetuar a extração de informação sobre novos documentos. Ao processo de extração pode ser necessário adicionar um processo de categorização das entidades. A forma de escrita de uma entidade pode variar de pessoa para pessoa como, por exemplo, '*Mercedes*' e '*Merc*'. O processo de categorização é responsável por identificar a entidade de interesse e todos os seus sinónimos e tratá-los como uma única entidade [24].

A extração dos aspetos de uma entidade, de um autor e de uma data em que a opinião foi efetuada, segue de uma forma geral os mesmos padrões de extração das entidades. A necessidade de categorização mantém-se, podendo ser aplicados os mesmos métodos para extração. A extração dos aspetos é um processo particularmente de maior dificuldade do que a extração dos restantes componentes. A análise de sentimentos ao nível dos aspetos vai ser abordada de forma mais completa posteriormente.

A extração das entidades, aspetos, autor e data da opinião ou combinações entre eles pode, em certos casos, ser ignorada dependendo dos requisitos impostos e da informação previamente conhecida. Por exemplo, na análise de um certo documento pode-se ter conhecimento prévio que o mesmo é referente a apenas uma entidade (a entidade de análise) e de qual o autor do texto e da sua data, bem como os aspetos que podem ser considerados irrelevantes para o nível de análise que se pretende. Neste exemplo, a extração destes componentes não é necessária. Um outro exemplo é dado por *Liu B.* [18] para a análise de sentimentos de opiniões provenientes de redes sociais e páginas web. Neste caso, o autor e a data das opiniões não são necessários extrair recorrendo a métodos de extração de informação como os apresentados acima uma vez que, normalmente, essa informação está contida de forma estruturada na página Web, sendo de fácil extração quando se recorre a técnicas de *Web Scraping*.

Extração e Classificação do Conteúdo Subjetivo

O processo de Extração e Classificação do Conteúdo Subjetivo tem a maior importância em análise de sentimentos. Dado que um documento pode ter vários tipos de informação, em análise de sentimentos, é necessário apenas a informação que contenha opiniões ou seja a informação subjetiva. O processo de classificação subjetiva é responsável por diferenciar o conteúdo objetivo do subjetivo e extrair este ultimo. Vários modelos foram já desenvolvidos baseados num grande numero de técnicas. De uma forma geral, todos eles têm um aspeto em comum: é desenvolvido especificamente para o contexto no qual ser aplicado. Isto deve-se à capacidade de várias palavras e frases que são representativas de subjetividade terem polaridades distintas em diferentes contextos.

A classificação de sentimentos é responsável por classificar o conteúdo subjetivo extraído no processo anterior. Os sentimentos podem ser classificados como positivos ou negativos. Em certos casos esta classificação binária não é suficiente e uma classificação nominal pode ser aplicada. Embora o processo de classificação subjetiva e classificação dos sentimentos seja abordada nos próximos capítulos com mais detalhe, aqui é apresentado um apanhado de algumas das técnicas tradicionais utilizadas como soluções para estes processos. Para classificação subjetiva foram utilizados métodos léxicos que fazem uso de palavras de sentimento anotadas com a respetiva polaridade [6, 22, 9]. Métodos de aprendizagem automática são também comuns. Nestes é utilizado um conjunto de documentos, manualmente anotados, que servem de treino para um classificador como *Naive-Bayes* para classificar a presença de subjetividade noutros documentos [36]. Outras estratégias estão relacionadas com a identificação de adjetivos em textos que são característicos da presença de subjetividade e ainda de combinações de partes do discurso. Uma outra estratégia foi utilizada por *Pang & Lee* [25] na qual foi utilizado *minimum cuts* para distinguir entre o conteúdo objetivo e subjetivo.

Várias técnicas foram desenvolvidas para a classificação de sentimentos que são baseadas em métodos de aprendizagem automática e em métodos léxicos. Recorrendo a um dicionário (método léxico) com informação de sinonímia e antonímia das palavras é possível criar um dicionário de sentimentos (dicionário de palavras de sentimento com respetiva polaridade) a partir de palavras *seed* cuja polaridade seja conhecida (bom e mau são duas palavras *seed* comuns). A forma do cálculo da polaridade difere de autor para autor. Por exemplo, *Kamps* [16] criou um grafo a partir de um dicionário de sinónimos e antónimos e utilizou o valor do caminho mais curto entre duas palavras para o calculo da polaridade. A utilização de um modelo de aprendizagem automática supervisionado foi implementado por *Pang et al* [27] para classificar positivamente e negativamente as opiniões de filmes onde foi utilizado *Naive-Bayes*, *Máxima Entropia* e *Support Vector Machines (SVM)* como métodos de classificação. Ainda para a classificação de opiniões de filmes, *Turney* [34] utilizou um modelo de aprendizagem

automática não-supervisionado. Ao contrário dos modelos supervisionados, estes modelos têm a vantagem de não necessitarem de grandes conjuntos de dados anotados para treino do modelo. O modelo desenvolvido utiliza partes do discurso para extrair conteúdo subjetivo e a classificação da polaridade desse conteúdo foi efetuada recorrendo ao calculo de *Pointwise Mutual Information (PMI)* para calcular a similaridade das palavras com as palavras *seed* "excelente" e "péssimo" utilizadas.

Resumo dos Resultados

Este último processo está relacionado com a forma como os sentimentos, que foram identificados anteriormente, são apresentados para observação e análise. O formato do resumo dos sentimentos pode variar consoante a aplicação e os requisitos necessários de análise. Um resumo baseado apenas num documento ou baseado em vários documentos são as formas tradicionais de resumir textos. Este último é a forma mais apropriada para aplicação em análise de sentimentos. Quando se pretende obter os sentimentos expressos sobre uma entidade, para que os sentimentos obtidos tenham um significado e importância relevantes, é necessário que sejam provenientes de vários textos, ou seja, que não sejam a reflexão das opiniões isoladas de um único autor mas de vários [26]. *Liu B.* [18] sustenta ainda que o resumo tradicional de textos baseados em vários documentos não é suficiente para uma boa apresentação de conteúdo útil. Para a definição de opinião anteriormente apresentada, o autor edifica que um resumo deve conter informação acerca da entidade e dos seus aspetos e ainda uma percentagem quantitativa associada a cada sentimento. Este tipo de informação é impossível de ser obtida através dos métodos tradicionais de resumo de textos uma vez que tipicamente os resumos obtidos por estes métodos são pequenos textos com a informação considerada mais relevante (resumo de apenas um documento) ou são textos estruturados com as diferenças entre vários documentos (resumo de vários documentos). Deve-se ainda enfatizar a importância da presença no resumo da percentagem quantitativa de cada sentimento. Uma percentagem pequena de sentimentos positivos é muito diferente de uma percentagem grande de sentimentos positivos sobre uma entidade (ou aspeto da mesma). Sem este tipo de informação torna-se difícil estabelecer uma noção da grandeza e importância dos sentimentos expressos.

Um resumo efetuado tendo por base a definição de opinião apresentada anteriormente denomina-se por resumo baseado em aspetos dado que a definição captura os aspetos da entidade em análise. Um resumo de sentimentos pode ter vários formatos como texto livre, texto formatado ou gráficos. A figura 2.2 mostra um tipo de resumo, baseado em aspetos, de uma câmara digital [14, 18].

O resumo da figura 2.2, em texto formatado, mostra claramente a entidade de análise e os

Digital Camera 1:

Aspect: **GENERAL**

Positive:	105	<individual review sentences>
Negative:	12	<individual review sentences>

Aspect: **Picture quality**

Positive:	95	<individual review sentences>
Negative:	10	<individual review sentences>

Aspect: **Battery life**

Positive:	50	<individual review sentences>
Negative:	9	<individual review sentences>

...

Figura 2.2: Resumo baseado em aspetos (exemplo) [18].

aspetos que foram opinados. O aspeto GERAL refere-se à câmara como um todo. Para cada aspeto é também apresentada informação da quantidade de opiniões positivas e negativas (equivalente à percentagem quantitativa). Uma outra forma de representar a informação da figura 2.2 é em formato gráfico, tal como é possível observar na figura 2.3 [19, 18]. Quando a análise de sentimentos é efetuada sobre mais do que uma entidade é possível apresentar graficamente o resumo das opiniões de cada uma em simultâneo para uma comparação rápida dos seus aspetos. A figura 2.4 [19, 18] mostra um gráfico deste tipo. Note-se que as entidades têm de pertencer ao mesmo domínio para ser possível compará-las – não se pode comparar, por exemplo, um automóvel com um telemóvel porque pertencem a domínios distintos.

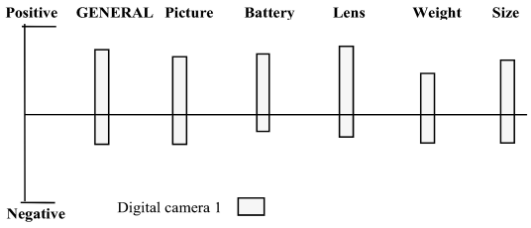


Figura 2.3: Resumo gráfico baseado em aspetos [18].

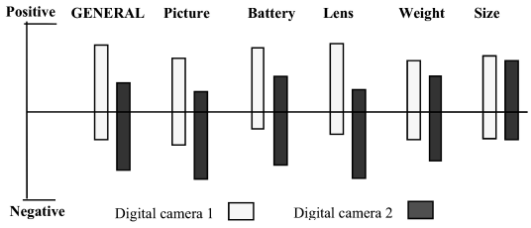


Figura 2.4: Resumo gráfico baseado em aspetos de duas entidades [18].

A informação apresentada nos resumo está naturalmente dependente da informação extraída nos processos anteriores. A forma, qualidade e tipo de dados que são recolhidos nesses processos influencia o tipo de informação e o modo como é apresentada no resumo. Pensando de um modo mais empresarial, ou seja, em soluções de análise de sentimentos mais completas (e complexas), a definição de opinião apresentada é ideal para servir de fonte de dados para um sistema de *data warehousing* seguido por um sistema OLAP que faz uso das capacidades de partição dos dados em várias perspetivas. Com a extração das datas das opiniões é possível agregar uma entidade em várias perspetivas e analisar as opiniões ao longo do tempo, mesmo para um sistema que não efetue o processo de extração e classificação de conteúdo subjetivo, efetuando apenas a extração das entidades, aspetos e datas das opiniões. Embora não seja um sistema de análise de sentimentos, este pode ser útil quando aplicado num sistema OLAP,

dado que permite observar ao longo do tempo os aspetos da entidade de análise que são mais relevantes para as pessoas (autores das opiniões) que para uma empresa e até para uma pessoa em particular pode ser útil [18].

2.3 Tipos de Opiniões

A forma como as pessoas expressam opiniões é virtualmente ilimitada. Como agravamento, as expressões utilizadas para proferir opiniões podem ter significados distintos em vários contextos. Até no mesmo contexto, uma mesma palavra pode exprimir opiniões diferentes para características distintas. O significado das opiniões expressas em textos é, portanto, complexo. Uma tentativa de simplificar a análise de opiniões é classifica-las em vários tipos como em [18] que sustenta quatro tipos de opiniões: regulares, comparativas, explícitas e implícitas.

2.3.1 Opiniões Regulares

- **Opiniões Regulares:** são o tipo de opiniões comuns, utilizadas mais frequentemente, e como tal são denominadas simplesmente opiniões.

As opiniões regulares podem ainda ser divididas em dois sub-tipos: diretas e indiretas.

- **Diretas:** são opiniões expressas diretamente sobre uma entidade ou um aspeto de uma entidade.

Exemplo 1: *"A qualidade de som é boa"*

O exemplo 1 expressa uma opinião regular direta. A maioria das técnicas desenvolvidas e da análise efetuada está direcionada para este tipo de opiniões dado que são muito mais simples de identificar e de determinar a entidade alvo da opinião.

- **Indiretas** – são opiniões expressas indiretamente sobre uma entidade ou aspeto de uma entidade baseada nos efeitos refletidos noutras entidades. Este tipo de opiniões são especialmente comuns no domínio médico.

Exemplo 2: *"A dor agravou-se depois de tomar o medicamento"*

O exemplo 2 é uma opinião regular indireta. Expõe um efeito indesejável da toma do medicamento sobre a dor sentida pela pessoa. Isto leva a que a opinião sobre o medicamento seja indiretamente negativa. Considere-se ainda o domínio médico, neste tipo de opiniões, é

necessário compreender se qualquer efeito desejável ou indesejável ocorre antes ou depois da toma de um medicamento.

Exemplo 3: *"Tomei o medicamento dado que tinha muitas dores"*

O exemplo 3 não expressa qualquer opinião sobre o medicamento uma vez que a "dor", que é algo negativo neste contexto, ocorreu antes de o tomar. Os exemplos acima (2 e 3) são característicos da dificuldade de análise deste tipo de opiniões.

2.3.2 Opiniões Comparativas

- **Opiniões Comparativas:** são opiniões que expressam similaridades ou diferenças entre duas ou mais entidades ou a preferência do autor da opinião sobre uma das entidades baseada em características comuns às duas.

As opiniões comparativas são, até um certo ponto, similares às opiniões regulares mas são também muito diferentes em relação à forma de escrita e significado das opiniões. Considere-se a opinião regular do exemplo 1 e a opinião comparativa abaixo:

Exemplo 4: *"A qualidade de som dos iPhones é melhor do que a dos Samsungs"*

Embora as duas opiniões sejam de certa forma similares, são também muito diferentes. Enquanto que a opinião regular indica que a qualidade de som é boa, a comparativa, neste exemplo, não indica que a qualidade de som de qualquer telemóvel é boa ou má mas limita-se simplesmente a compara-la.

Exemplo 5: *"Os iPhones são um centímetro mais baixos do que os Samsungs"*

As opiniões comparativas podem ainda não exprimir qualquer sentimento expondo apenas informação objetiva da comparação de duas ou mais entidades. O exemplo 5 não expressa qualquer sentimento enquanto que no exemplo 4 existe um sentimento explícito da preferência do autor em relação à qualidade de som dos dois telemóveis.

2.3.3 Opiniões Explícitas e Implícitas

- **Opiniões Explícitas** – são opiniões diretas, constituídas por opiniões regulares (exemplo 1) ou comparativas (exemplo 4).

- **Opiniões Implícitas** – são opiniões expressas em conteúdos objetivos. O sentimento associado está implícito na exposição dos factos e normalmente expressa algo desejável ou indesejável acerca da entidade ou das suas características. São expressas implicitamente em opiniões regulares (exemplo 6) ou comparativas (exemplo 7).

Exemplo 6: *"Comprei o telemóvel ontem e já não funciona"*

Exemplo 7: *"A duração da bateria dos iPhones é melhor do que os Samsungs"*

Enquanto que as opiniões explícitas, na maioria das vezes, fazem uso de palavras de sentimento, que são características de presença de sentimentos, as opiniões implícitas, cuja natureza é maioritariamente objetiva, não faz uso de tais palavras de sentimento, expondo apenas factos acerca da entidade. Isto leva a que as opiniões explícitas sejam muito mais fáceis de identificar e classificar. A utilização de palavras de sentimento e técnicas que fazem uso destas palavras para identificar sentimentos em textos vão ser apresentadas nos próximos capítulos.

2.4 Sentimentos em Textos

O processo de extração de sentimentos em conteúdos textuais levanta a questão de como são expressos os sentimentos em textos. Independentemente do tipo de texto em análise, *reviews*, blogs, redes sociais, entre outros, efetuar a análise da forma como os sentimentos são expressos é o primeiro passo a tomar. Diferentes tipos de textos possuem características específicas de escrita e de expressão do conteúdo, quer o tipo seja objetivo ou de opinião. Enquanto que, para textos provenientes de fontes como o *twitter*, que são predominantemente curtos e com linguagem descuidada, qualquer opinião expressa sobre uma entidade é efetuada de forma rápida e clara, para textos provenientes de blogs, característicos pela longevidade e pela qualidade de linguagem, as opiniões presentes podem abranger várias entidades, podendo-se encontrar dispersas ao longo do texto e podem ser expressas em formas complexas como comparações, ironia ou implicitamente, o que leva à necessidade de uma pré-compreensão do tema em discussão [26]. Como é possível perceber, a resolução de comparações, ironia e diferentes tipos de opiniões são alguns dos muitos obstáculos a serem contornados, uma vez que dificultam a identificação e a classificação da polaridade dos sentimentos. Estes e outros problemas vão ser analisados posteriormente.

2.4.1 Identificação de Sentimentos

Até agora tem-se denominado as fontes de informação como textos mas num sentido mais abrangente as mesmas são denominadas documentos. Documentos não são nada mais do que a junção de vários textos provenientes de uma ou mais fontes com informação acerca da área de análise. Note-se que uma simples frase é também considerada um documento [21].

A forma como os conteúdos textuais são apresentados varia de acordo com a área de análise a que se destinam. Todavia, o conteúdo apenas pode ser apenas de dois tipos gerais, nomeadamente:

- **Conteúdo objetivo** – constituído por informação factual, ou seja, apresenta um conjunto de factos sobre a entidade que trata apoiando-se nas suas características para fundamentar a informação.
- **Conteúdo subjetivo** – apresenta informação de opinião, sentimento ou pontos de vista e que é expressa sobre as entidades de análise e suas características.

A presença de sentimentos em textos é identificada, normalmente, em conteúdos subjetivos. Existe aqui a necessidade de salientar a palavra normalmente utilizada atrás. Embora a presença de subjetividade seja característica de opiniões, alguns conteúdos objetivos podem também exprimir sentimentos. Considere-se:

Exemplo 1: *"Esta máquina de lavar utiliza muita água"*

No exemplo acima, apenas é constatado um facto acerca do desperdício de recursos pela máquina de lavar. Contudo, tendo em conta o contexto, existe um sentimento negativo associado à exposição dos factos uma vez que o consumo excessivo de recursos como a água, neste contexto, é uma característica não desejável [18]. Este é um problema da análise de sentimentos que está relacionado com a presença de sentimentos implícitos dependentes do contexto e que vai ser discutido posteriormente.

Os sentimentos são, então, caracteristicamente subjetivos, propriedade que é de grande importância no desenvolvimento de uma aplicação. Esta característica torna importante a obtenção de opiniões provenientes de várias fontes e não de apenas uma, porque neste último caso, a opinião acerca da entidade, é influenciada por apenas uma fonte o que não é suficientemente forte (na maioria dos casos) para tirar conclusões acerca da entidade.

A identificação de sentimentos em textos preocupa-se em distinguir entre conteúdos objetivos e subjetivos. Este processo denomina-se por classificação subjetiva.

- **Classificação Subjetiva** – processo responsável por identificar, de entre o conteúdo de um documento, aquele que é subjetivo e lidar com vários problemas como a presença de subjetividade em conteúdos objetivos como foi demonstrado no exemplo 1.

A classificação subjetiva é um processo que pode ser aplicado em várias áreas de análise para além da análise de sentimentos e tem vindo a ser desenvolvida de forma cada vez mais eficiente ao longo dos anos. Contudo, é considerado por muitos autores que ainda se está longe de conseguir uma solução com precisão verdadeiramente satisfatória para ser integrada em aplicações cuja precisão dos resultados seja crucial. De facto, a correta classificação de subjetividade é em muitos casos mais improbo do que a subsequente classificação de polaridade no contexto de análise de sentimentos. Qualquer avanço na capacidade de classificação de subjetividade leva a um impacto positivo no processo de análise de sentimentos [28].

A subjetividade pode ser expressa de várias formas, em vários tipos de texto recorrendo a múltiplos formatos de expressões. As palavras e o seu sentido são os pilares que sustentam a presença de subjetividade. Em casos mais complexos, como em textos de notícias, políticos e debates, a análise das palavras do texto não é suficiente para identificar a presença de subjetividade com uma precisão aceitável. Nestes casos é necessário um maior conhecimento acerca do contexto a que o texto se refere e por vezes o titular das opiniões expostas [20].

As várias técnicas utilizadas para classificação subjetiva como aprendizagem automática [36, 25] ou léxicos [6, 22, 9] têm como base de análise as palavras, que são os maiores indicadores de sentimentos. Estas palavras são designadas por palavras de sentimento e são comummente utilizadas para expressar sentimentos – na figura 2.5 podemos ver algumas destas palavras.

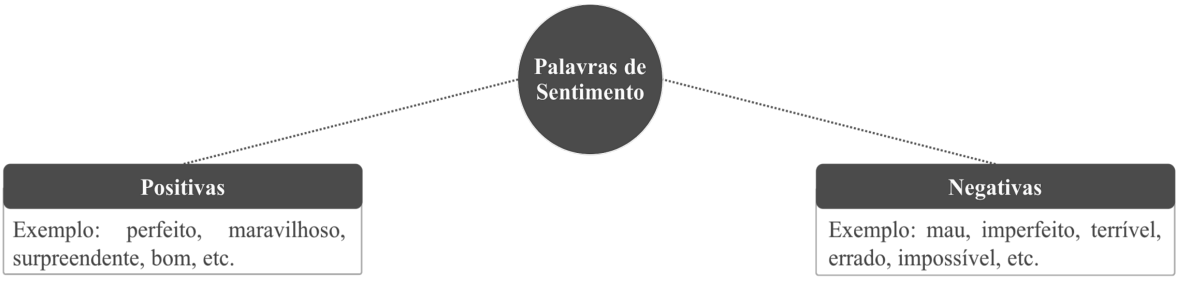


Figura 2.5: Exemplo de palavras de sentimento.

Como existem opiniões regulares e comparativas é de esperar que as palavras de sentimento também possam ser categorizadas desta forma. Ao contrário das palavras de sentimento regulares, as palavras de sentimento comparativas não expressam uma opinião regular sobre uma entidade, mas sim uma opinião comparativa sobre duas ou mais entidades. A figura acima apresenta apenas palavras de sentimento regulares. A presença de palavras de sentimento positivas implica um sentimento positivo em relação à entidade de análise, enquanto que as palavras de sentimento negativas implicam, naturalmente, um sentimento negativo. Para além

de palavras de sentimento também podem ser utilizadas frases predefinidas cujo sentimento seja conhecido:

Exemplo 2: *"Esta máquina custa os olhos da cara"*

O exemplo 2 apresenta uma frase popular com um sentimento negativo associado, embora não possua qualquer palavra de sentimento positiva ou negativa. Este tipo de casos é de extrema dificuldade para qualquer sistema de classificação subjetiva reconhecer como subjetivo e, consequentemente, o modelo de análise de sentimentos falha em atribuir qualquer sentimento. Este tipo de frases devem então ser identificadas previamente através da análise do contexto sobre o qual a solução vai ser aplicada. O conjunto de palavras e frases de sentimento, com respetivas polaridades, denomina-se um dicionário de sentimentos.

Existem muitas outras técnicas utilizadas por vários autores para identificar subjetividade em textos para além da utilização de palavras de sentimento. A marcação de partes do discurso *(POS tagging)* é uma das técnicas mais utilizadas e está relacionada com o reconhecimento de palavras de sentimento mas vai mais longe do que simplesmente identificar este tipo de palavras, uma vez que consegue por si só classificar subjetivamente as partes constituintes de um documento. A utilização de palavras de discurso assenta na classificação das classes gramaticais das palavras de um documento. Considere-se as palavras de sentimento apresentadas anteriormente na figura 2.5. Todas elas têm algo em comum, a classe gramatical a que pertencem é a mesma: adjetivos. Existe uma grande correlação entre adjetivos e a presença de subjetividade em textos [13]. Esta relação levou a que a presença de adjetivos seja utilizada amplamente para encontrar palavras de sentimento e determinar as suas polaridades, mas também para encontrar características das entidades presentes no texto visto que um adjetivo classifica, na maioria das vezes, um aspeto que se encontra na sua proximidade.

Durante algum tempo, a aplicação de palavras do discurso para identificar a presença de subjetividade baseava-se apenas na utilização de adjetivos. Contudo, o facto de que adjetivos são bons para determinar a presença de subjetividade não significa que outras partes do discurso não possam ser utilizadas. Nomes, verbos, advérbios e substantivos são também bons indicadores de subjetividade quando aplicados em conjunto (não necessariamente todos em conjunto). De facto, existe um grande número de partes do discurso que podem ser usadas para identificar subjetividade em textos. Na tabela 2.1 podemos ver aquele que é considerado o conjunto de partes do discurso mais completo (e complexo) aplicado na literatura: *Penn Treebank POS Tags* [30].

A conjunção de partes do discurso pode então ser utilizada para classificação de subjetividade e posterior classificação de polaridade dos sentimentos. Um exemplo deste tipo de conjunção foi desenvolvido por *Turney, 2002* [34]. O método desenvolvido utiliza padrões de partes do

discurso que são mais propícios para exprimir sentimentos. A tabela 2.2 mostra os padrões utilizados. O método analisa conjuntos de três palavras/partes do discurso para extrair apenas duas (bigramas).

Exemplo 3: *"This piano produces beautiful sounds"*

Tag	Description	Tag	Description
CC	Coordinating conjunction	PRP$S	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential there	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	to
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3d person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wb-adverb

Tabela 2.1: *Penn Treebank POS tags* – partes do discurso.

A aplicação dos padrões na frase do exemplo 3 [18] faz com que a mesma seja classificada como subjetiva, dado que satisfaz o primeiro padrão da tabela 2.2, enquanto que as palavras *"beautiful sounds"* são extraídas para posterior classificação da polaridade do sentimento expresso. Este é um bom exemplo para demonstrar a capacidade da utilização de partes do discurso para determinar tanto subjetividade como características (ou aspetos) da entidade a que os sentimentos são dirigidos. Como referido anteriormente – *"sounds"* é a característica do piano a que o sentimento se dirige. Note-se que os padrões da tabela 2.2 foram desenvolvidos tendo como alvo de análise a linguagem inglesa, a aplicação noutras linguagens necessita de uma alteração apropriada aos padrões de forma a refletir a linguagem de análise.

Todos estes métodos podem, também, ser utilizados como partes integrantes do processo de treino em métodos de aprendizagem automática. Estes métodos podem ser utilizados para classificação subjetiva pela utilização de padrões anotados ou não que podem fazer uso de conjunções de palavras de sentimento como em [37].

	First word	Second word	Third word (not extracted)
1.	JJ	NN or NNS	anything
2.	RB, RBR, or RBS	JJ	not NN nor NNS
3.	JJ	JJ	not NN nor NNS
4.	NN or NNS	JJ	not NN nor NNS
5.	RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything

Tabela 2.2: Padrões de partes do discurso para extração de duas palavras.

2.5 A Classificação de Sentimentos

Após a identificação dos sentimentos presentes nos textos, efetuada pelo processo de classificação subjetiva, é possível recorrer a um conjunto de métodos de classificação da polaridade do conteúdo subjetivo identificado.

- **Classificação de Sentimentos** – processo de classificação do conteúdo subjetivo em positivo ou negativo.

A classificação dos sentimentos não é restrita a positivos e negativos. Dependendo das necessidades que se procuram resolver, podem ser utilizadas outras formas de classificação como sentimentos neutros ou classificações nominais. Pode-se utilizar ainda um nível de especificidade maior na classificação dos sentimentos associando um grau de intensidade a cada sentimento positivo ou negativo. Esta última forma de classificação permite diferenciar sentimentos de uma mesma classe onde um sentimento positivo de intensidade alta é diferente de um sentimento positivo de intensidade baixa. Este nível de granularidade oferecido por este método pode ser bastante útil quando se pretende obter o maior nível de precisão possível mas é também o método de maior dificuldade de implementação e consequente obtenção de boa precisão. Vários modelos para classificação de sentimentos foram desenvolvidos ao longo dos anos e aplicados na literatura. A figura 2.6, adaptada de [21], apresenta alguns desses modelos.

Os modelos de classificação de sentimentos podem ser divididos em modelos de aprendizagem automática e modelos léxicos. Os modelos de aprendizagem automática podem ser supervisionados ou não-supervisionados. Os modelos supervisionados podem ainda ser classificados em vários tipos como é possível observar na figura 2.6 que também apresenta os vários tipos dos modelos léxicos. Inicialmente eram utilizados modelos léxicos baseados em dicionários. Porém, com o desenvolvimento do processamento de linguagem natural foram desenvolvidos modelos de aprendizagem automática com precisões cada vez mais elevadas, que rapidamente se tornaram nos modelos mais utilizados para classificação de sentimentos.

Na literatura podemos ainda encontrar referência a modelos híbridos. Estes modelos tiram

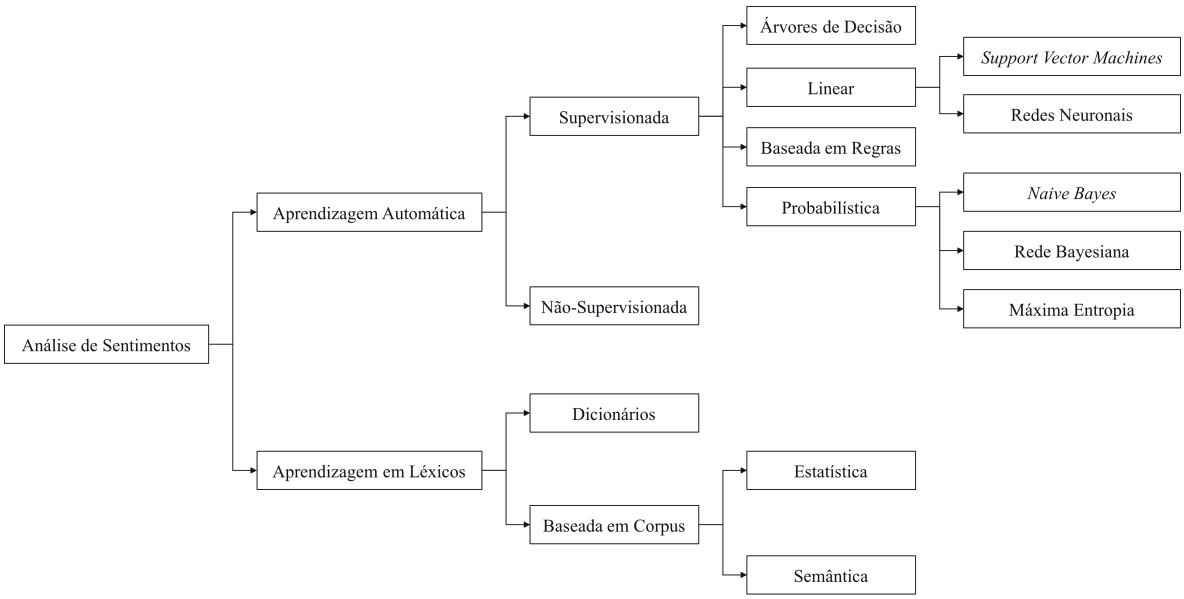


Figura 2.6: Modelos de análise de sentimentos [21].

partido de características dos modelos léxicos e de aprendizagem automática. A combinação de modelos não-supervisionados de aprendizagem automática com modelos léxicos baseados em dicionários é muito comum na construção de um modelo híbrido [4].

De entre os modelos de classificação de sentimentos apresentados na figura 2.6 existem alguns que são mais frequentemente utilizados. Nos modelos de aprendizagem automática, *Naive Bayes*, *Support Vector Machines* e modelos não-supervisionados são os mais utilizados, enquanto que nos modelos léxicos aqueles que são baseados em dicionários são os mais utilizados. Os modelos híbridos são igualmente utilizados. Em muitas soluções conseguem obter mesmo um nível de precisão consideravelmente superior aos restantes modelos. Antes de apresentar informação mais detalhada acerca de cada um dos modelos é necessário realçar que nenhum deles é melhor do que outro. Uma boa precisão resultante da utilização de um dado modelo sobre um problema não significa que a utilização noutro problema consiga obter o mesmo nível de precisão. O modelo utilizado deve então ser desenvolvido especificamente para o contexto de análise referente a um problema específico.

2.5.1 Dicionários

Os dicionários são um modelo léxico muito utilizado no processo de classificação de sentimentos. Contudo, a utilização individual de dicionários para esta tarefa não é muito comum devido a um conjunto de problemas inerentes aos mesmos e também porque geralmente a utilização de aprendizagem automática consegue obter resultados superiores. Os dicionários são então,

muito utilizados, não de forma individual, mas em conjunto com modelos de aprendizagem automática não-supervisionados cuja conjunção forma um processo híbrido.

Palavras de sentimento como as da figura 2.5, são essenciais para identificação de subjetividade e classificação de sentimentos. Um dicionário de sentimentos é constituído por um conjunto de palavras de sentimento anotadas com a sua respetiva polaridade (pares palavra/polaridade). Podem ainda ser considerados outros dicionários de sentimentos mais completos, que para além de possuírem informação da polaridade de cada palavra, possuem também informação acerca da intensidade da polaridade, permitindo diferenciar de forma muito mais precisa entre polaridades idênticas e distintas.

Implementação

Existem vários métodos para a construção de um dicionário de sentimentos, desde o método mais simples [14], representado na figura 2.7, até métodos mais complexos que conseguem ultrapassar, até um certo ponto, alguns dos problemas característicos dos dicionários.

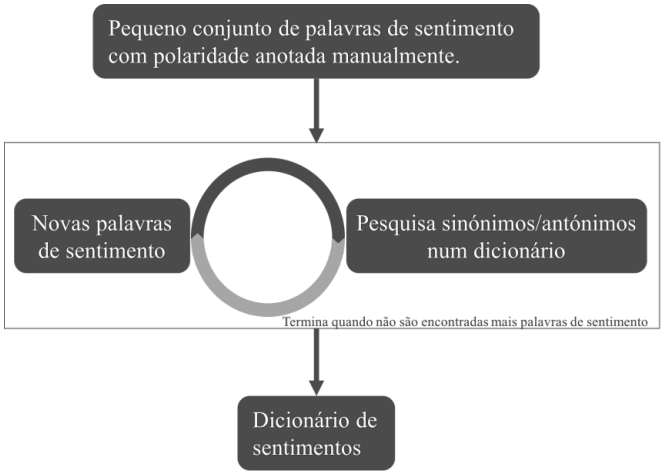


Figura 2.7: Construção de um dicionário de sentimentos.

O método mais simples utilizado na literatura para construção de um dicionário de sentimentos começa com um pequeno conjunto de palavras de sentimento anotadas com a sua polaridade. Este conjunto de palavras é denominado de palavras *seed* e é constituído no mínimo por duas palavras, uma com polaridade positiva e outra com polaridade negativa. Com a utilização de um dicionário que possua informação de sinonímia e antonímia das palavras como *WordNet* (apenas para inglês), são procurados sinónimos e antónimos das palavras *seed* que originam novas palavras *seed*. O processo repete-se até que não existam mais palavras por processar, ou seja, até que todas as palavras de sentimento com respetivas polaridades tenham sido construídas. Todo o processo assenta na hipótese de que sinónimos de palavras de sentimento

positivas são também positivos e antónimos são negativos e o mesmo raciocínio é aplicado para as palavras de sentimento negativas. O resultado final é um dicionário de palavras de sentimento com as respetivas polaridades.

Quando o problema de análise é referente a um contexto específico, o dicionário deve ainda possuir informação da polaridade de frases comuns desse contexto que podem possuir um sentimento mas sem utilizarem palavras de sentimento, como acontece no exemplo 1 da secção 2.4. Este tipo de frases devem, então, ser identificadas à priori e acrescentadas manualmente ao dicionário ou sub-processo da sua identificação, das mesmas com a polaridade que expressam.

A utilização de dicionários para classificação de sentimentos passa, na forma mais básica, por consulta direta ao dicionário de palavras de sentimento de cada palavra identificada como subjetiva no documento de análise. É, então, imperativa a necessidade de construção de um dicionário o mais completo possível para que não ocorram casos de procura de uma palavra que não se encontre no dicionário.

Problemas

A utilização de um dicionário de sentimentos é um modelo simples e útil de classificação de sentimentos mas possui alguns problemas que fazem com que a utilização individual não seja o melhor método possível. *Liu B.* [18] refere que a utilização de um dicionário é essencial, mas não é suficiente para obter uma boa precisão. O autor refere ainda alguns dos problemas da utilização individual de dicionários, nomeadamente:

- **Contexto** – Uma palavra de sentimento pode ter polaridades distintas dependendo do contexto no qual é aplicada, como mostram os exemplos:

  Exemplo 1: *"O altifalante do telemóvel é silencioso"*

  Exemplo 2: *"O carro é silencioso"*

  Nos exemplos acima, a palavra de sentimento "silencioso"exprime sentimentos opostos devido ao contexto no qual está inserido. O primeiro exemplo exprime, normalmente, um sentimento negativo enquanto que o segundo exemplo um sentimento positivo.

- **Uma frase que tenha palavras de sentimento pode não exprimir qualquer sentimento** – A presença de palavras conhecidas de sentimento nem sempre implica a presença de sentimentos. Isto é comum em vários tipos de frases e mais frequente

em frases interrogativas e condicionais. Vejamos os seguintes exemplos:

Exemplo 3: *"Qual a melhor câmara da Sony?"*

Exemplo 4: *"Se encontrar uma boa câmara na loja, vou compra-la"*

Os exemplos utilizam as palavras de sentimento "melhor" e "boa" respetivamente, mas nenhuma exprime qualquer sentimento sobre uma câmara.

O problema agrava-se quando se constata que existem frases interrogativas e condicionais que podem exprimir sentimentos como no exemplo 5 abaixo, onde existe um sentimento positivo sobre a marca de automóvel referida, proveniente da palavra de sentimento "bom".

Exemplo 5: *"Se queres um bom carro, compra um Rolls Royce"*

- **Sarcasmo** – Frases sarcásticas, que tenham ou não palavras de sentimento, são muito difíceis de resolver.

  Exemplo 6: *"Que carro maravilhoso, deixou de funcionar em dois dias"*

- **Frases que não tenham palavras de sentimento podem exprimir sentimentos** – Comum em frases objetivas como no exemplo 1 da secção 2.4.

2.5.2 Modelos Supervisionados

Os modelos supervisionados (e não-supervisionados) de aprendizagem automática têm sido aplicados em grande escala em toda a variedade de problemas de NLP e, consequentemente, todo o conhecimento ganho auxiliou na aplicação dessas técnicas em análise de sentimentos. Estes modelos necessitam de grandes quantidades de dados anotados, que são usados para treino do modelo. A quantidade e qualidade dos dados de treino é determinante para a precisão do modelo. Isto representa o primeiro obstáculo da aplicação destes modelos em análise de sentimentos, uma vez que, devido à falta de maturidade da área (apenas a partir de 2000 é que a área ganhou interesse), não existem grandes conjuntos de dados anotados para vários contextos nem formas eficientes de unificação dos contextos para que apenas um conjunto de anotações sejam aplicadas com sucesso em qualquer contexto. Para além disto, os poucos dados de treino disponíveis publicamente têm como alvo a língua inglesa, o que torna difícil a sua aplicação noutras línguas.

Dados de Treino

Existem vários algoritmos de modelos supervisionados, alguns deles já apresentados na figura 2.6 atrás, cuja precisão dos resultados varia entre eles consoante o problema. No caso de análise de sentimentos, a escolha do algoritmo é menos importante do que o tratamento dos dados. Este tratamento permite efetuar uma seleção de um conjunto de características textuais, relevantes para análise de sentimentos, que representam sentimentalmente um texto e que são utilizadas para treino do modelo [33]. A mesma seleção de características deve ser efetuada em novos documentos de análise para que seja possível aplicar o modelo sobre os mesmos. Algumas das características textuais utilizadas para treino de modelos supervisionados em análise de sentimentos são:

- **Palavras e suas frequências** – também denominado por *"bag of words"*, é um dos métodos mais utilizados. Utiliza o conjunto de palavras presentes no texto com a respetiva frequência (número de vezes que ocorrem). Este método é baseado em unigramas, podendo, contudo, ser utilizados bigramas (conjuntos de duas palavras e frequências), trigramas e assim sucessivamente, dependendo de qual consegue obter a melhor precisão.

  Em acréscimo da utilização de unigramas ou conjuntos de n-gramas de palavras, podem ser utilizadas técnicas de atribuição de pesos às palavras como *Term Frequency-Inverse Document Frequency (TF-IDF)*. De uma forma simples, TF-IDF pode ser separado em duas partes: TF e IDF. A parte TF (inter-documento) é responsável por aumentar o peso das palavras que ocorrem mais frequentemente num documento, sendo escalonado de forma logarítmica para que o peso de uma palavra não seja exageradamente alto (uma palavra que ocorra 20 vezes num documento muito dificilmente possui importância 20 vezes superior à ocorrência de apenas uma vez). A parte IDF (intra-documento) reduz o peso das palavras que ocorrem mais frequentemente em todos os documentos de treino e aumenta o peso das palavras que ocorrem menos frequentemente nos documentos de treino. Desta forma, palavras raras, que muito provavelmente são importantes para o contexto, conseguem um tratamento especial aumentando o seu peso, enquanto que palavras que ocorrem mais frequentemente em todos os documentos (como "de", "a", "o", "em", etc) têm o seu peso reduzido dado que não trazem qualquer tipo de informação relevante.

  A título de exemplo, considere-se a seguinte frase, de José Saramago: *"Não tenhamos pressa, mas não percamos tempo"*. Os exemplos 7 e 8 mostram o resultado de aplicação

de unigramas e bigramas, respetivamente, sobre a frase.

Exemplo 7: *Não | tenhamos | pressa | , | mas | não | percamos | tempo*

Exemplo 8: *Não tenhamos | tenhamos pressa | pressa , | , mas | mas não | não percamos | percamos tempo*

Apesar deste método ser um dos mais utilizados e conseguir uma boa precisão, *Pang et al* [27] mostrou que a utilização das frequências nem sempre consegue obter os melhores resultados na classificação de sentimentos. Os autores utilizaram apenas a presença das palavras de forma binária (1 caso a palavra ocorra no texto e 0 caso contrário). Para o conjunto de dados utilizados, a precisão deste modelo foi superior à utilização de frequências. Em [26] é referido que isto pode ser indicativo de que para a classificação de um tópico, a frequência das palavras talvez seja a maior indicação do mesmo enquanto que para a classificação de sentimentos a frequência com que as palavras ocorrem talvez não seja o melhor identificador em comparação com a simples ocorrência de palavras.

- **Partes do Discurso (POS)** – A utilização de partes do discurso já foi apresentada atrás como um possível método de classificação subjetiva. A utilização em classificação de sentimentos segue basicamente os mesmos princípios. Dado que adjetivos e outras partes do discurso (tabela 2.1) e conjunções entre elas (tabela 2.2) são grandes indicadores de subjetividade, a presença de certos padrões podem ser representativos de certos sentimentos. Esta correspondência pode ser utilizada para treino de um modelo supervisionado de aprendizagem automática que pode depois ser aplicado para classificação de padrões em documentos de análise.

- **Modificadores de Sentimentos (*sentiment shifters*)** – A polaridade de um sentimento pode ser facilmente modificada pela utilização de modificadores de sentimentos. As palavras de negação são os modificadores de sentimentos mais comuns e são representadas por advérbios de negação. A palavra de negação mais comum em português é "não" mas muitas outras podem ser utilizadas bem como frases específicas de contexto.

  A identificação de negações é um dos processos mais difíceis e cruciais em análise de sentimentos. Pode ser implementada como um sub-processo que faz uso de técnicas léxicas ou de aprendizagem automática. Quando uma negação é identificada, a forma generalizada de introdução dessa negação no processo de classificação é acrescentar "NAO" às palavras que estão perto da negação como no exemplo 9 ou a todas as palavras depois da negação até certos sinais de pontuação, palavras de sentimento ou

negações como no exemplo 10.

Exemplo 9: *"Eu não gostei-NAO do filme"*

Exemplo 10: *"Eu não gostei-NAO do-NAO filme-NAO nem dos-NAO atores-NAO. O enredo era bom"*

Contudo, a identificação de negações não se pode resumir à procura de palavras de negação porque estas palavras nem sempre alteram a polaridade dos sentimentos como no exemplo 11. Este tipo de caso deve ser reconhecido à priori.

Exemplo 11: *"Não só ... mas também"*

Após a identificação das negações podem ser aplicados os métodos de palavras e suas frequências ou partes do discurso atrás apresentadas para efetuar o treino do modelo.

Vantagens & Desvantagens

Como em qualquer outro modelo, a utilização de modelos supervisionados para classificação de sentimentos contém algumas vantagens e desvantagens. A principal vantagem é a capacidade de adaptação sobre o contexto para o qual é desenvolvido. Os dados de treino quanto mais representativos forem do contexto de análise melhor vai ser a precisão dos resultados nesse contexto. As principais desvantagens são:

- Necessita de um grande conjunto de dados de treino anotados.
- Nem sempre existem dados de treino anotados. Os processos envolventes na análise de sentimentos são dependentes do contexto que faz com que os dados anotados que existem sejam específicos para o contexto sobre o qual foram anotados tornando difícil a aplicação noutros contextos. A criação de dados anotados pode ser efetuada manualmente, mas devido à grande quantidade necessária torna a anotação muito demorada. Podem também ser aplicados modelos de aprendizagem automática para anotação automática dos dados. Embora a utilização destes modelos automatize o processo de anotação, os resultados estão suscetíveis a erros sendo necessário uma revisão manual dos mesmos.
- A quantidade e qualidade dos dados de treino do modelo é crucial para uma boa precisão dos resultados.
- Embora um modelo desenvolvido tenha a vantagem de ser aplicável ao contexto no qual vai ser aplicado, esse mesmo modelo não é aplicável a vários contextos.

2.5.3 Modelos Não-Supervisionados

A aplicação de modelos não-supervisionados é vastamente referida na literatura. De facto, consegue-se perceber os motivos que levam à escolha de modelos não-supervisionados em detrimento de modelos supervisionados e dicionários, dado que estes modelos conseguem, de certa forma, ultrapassar alguns dos problemas dos restantes modelos. Os modelos não-supervisionados fazem uso da conjunção de vários recursos dos modelos supervisionados, dicionários, modificadores de sentimentos, tratamento de palavras de contraste (*but-clauses*) e outros dependendo do contexto. Ao efetuar a conjunção de todos estes recursos, estes modelos permitem que problemas característicos dos restantes modelos, anteriormente apresentados, sejam ultrapassados, principalmente a não necessidade de um conjunto de dados anotados para treino do modelo.

Deve-se fazer uma nota de advertência em relação à utilização livre na literatura do termo "não-supervisionado" referente a um dado modelo. Alguns autores como *Liu B.* [18] utilizam esta nomenclatura para caracterizar qualquer tipo de modelo que não seja supervisionado. Assim, um modelo baseado em dicionários pode ser considerado também um modelo não-supervisionado, embora aqui tenha sido apresentado como um modelo baseado em léxicos.

Modelos não-supervisionados utilizam, então, métodos de extração de características de sentimentos para conseguir classificar os sentimentos presentes nos documentos de análise. Um dos métodos mais adotados é a utilização de partes do discurso para deteção de sentimentos (classificação subjetiva) como apresentado anteriormente. Dada a grande variedade de recursos e de métodos que podem ser aplicados nos modelos não-supervisionados, a melhor forma de demonstrar a forma como são utilizados é exemplificando. Considere-se, assim, o seguinte modelo não-supervisionado, implementado em [34], constituído por três passos principais:

1. Identificação de duas palavras consecutivas com a utilização de partes do discurso, de acordo com os padrões já apresentados na tabela 2.2, que são característicos da presença de subjetividade. Nesta fase são extraídas as frases nas quais sejam identificados esses padrões.

2. A polaridade dos sentimentos das frases extraídas é obtida pelo calculo de PMI de cada palavra de sentimento das frases em relação às palavras *seed* "excelente" e "mau". Uma palavra de sentimento tem polaridade positiva o quanto mais associada for com a palavra "excelente" e tem polaridade negativa o quanto mais associada for com a palavra "mau".

3. A polaridade de um documento de análise é calculada pela média dos valores da polaridade obtida em cada frase do documento. A polaridade do documento é positiva se a média for positiva e é negativa caso contrário.

Este é apenas um exemplo de uma implementação de modelos não-supervisionados. Muitos outros foram desenvolvidos com propósitos distintos, como aconteceu em [29], em que foi desenvolvido um modelo não-supervisionado para classificação de sentimentos que não seja dependente do contexto, ou seja, que possa ser utilizado para determinar a polaridade de qualquer documento independentemente do contexto em que se insere. Na próxima secção será também apresentado outro exemplo de uma possível implementação de um modelo não-supervisionado ao nível dos aspeto de classificação de sentimentos.

2.6 Níveis de Classificação

Até este momento, não foi especificado qualquer nível de análise sobre os documentos. Um documento de análise pode ser visto como um conjunto de grandes textos ou aglomeração de vários documentos bem como apenas uma simples frase. Esta característica diversificada de um documento permite, então, definir com mais precisão o alvo de análise de um documento: existe uma generalidade de adoção pelos autores, como em [21] e [18], de três níveis de classificação sobre os documentos:

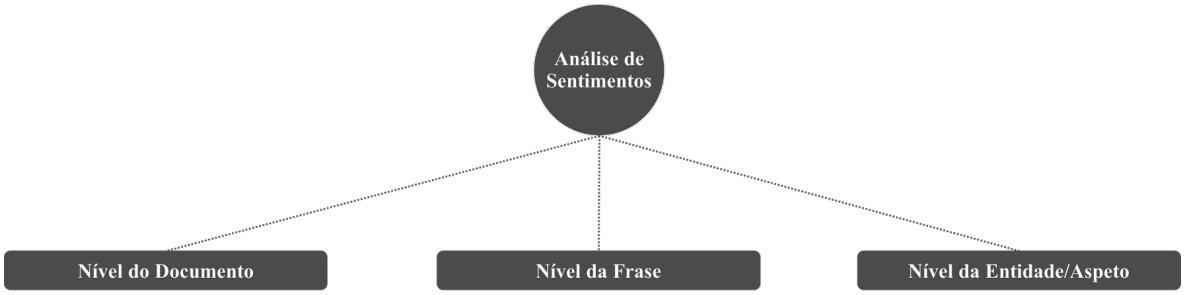


Figura 2.8: Níveis de classificação.

Os níveis de análise apresentados na figura 2.8 permitem alterar o nível e alvo sobre o qual o processo de análise de sentimentos é aplicado. Qualquer que seja o nível de análise, os métodos de classificação subjetiva e de classificação de sentimentos atrás apresentados mantêm-se, embora alguns deles sejam mais apropriados para certos níveis de classificação conseguindo melhores resultados.

2.6.1 Nível do Documento

Este nível de classificação representa o tipo de análise até agora apresentada. Todo o documento é utilizado como uma unidade base de informação com o objetivo de classificar a

opinião geral do documento como positiva ou negativa. Por exemplo, dado um documento de opinião acerca de um produto, este nível de classificação determina se a opinião geral sobre o produto é positiva ou negativa.

Liu B. [18] enfatizou uma assunção acerca deste nível de classificação: o documento de análise apresenta opiniões provenientes de apenas uma pessoa (um autor) sobre apenas uma entidade. Esta assunção é necessária para este nível, dado que se o documento exprimir opiniões sobre mais que uma entidade, então os sentimentos expressos sobre cada entidade podem ser distintos e caso exista mais que um autor, podem ter sentimentos distintos, fazendo com que não seja prático classificar todo o documento com apenas uma polaridade.

Embora qualquer um dos modelos de classificação apresentados possam ser utilizados neste nível de análise, modelos supervisionados são os mais utilizados dada a natureza binária dos sentimentos (positivos ou negativos). No caso da aplicação de uma escala nominal, modelos de regressão devem ser aplicados. A utilização deste nível de classificação deve ter em conta algumas desvantagens dependentes da aplicação, em particular:

- A assunção acima apresentada não permite que seja facilmente aplicado a, por exemplo, forums, blogs ou outros textos de opinião que sejam longos dado que em muitos casos, estes tipos de textos são característicos por apresentar opiniões acerca de várias entidades com comparações entre as mesmas.

- O nível do documento pode não ser suficientemente específico sendo, em muitos casos, necessário saber quais os aspetos da entidade (ou entidades) que são tidos como positivos e negativos.

2.6.2 Nível da Frase

Este nível de classificação tem um alvo mais específico do que o nível anterior. O nível da frase divide o documento de análise em frases individuais e é sobre cada uma dessas frases que o processo de análise de sentimentos é aplicado.

Tal como acontece no nível do documento, *Liu B.* [18], refere uma assunção que é comummente adotada: cada frase contém apenas uma opinião proveniente de apenas um autor. A presença de apenas uma opinião por frase serve como uma forma de simplificação dado que isso nem sempre acontece como é possível ver no exemplo 1.

Exemplo 1: *"A qualidade de imagem desta câmara é boa mas o visor é demasiado pequeno"*

O exemplo acima não respeita a assunção referida, dado que possui duas opiniões e cada uma

delas com sentimentos distintos que faz com que seja complicado atribuir um único sentimento preciso à frase. Contudo, esta assunção não limita a utilização deste nível de classificação sobre frases como as do exemplo acima. *Hu e Liu* [14] desenvolveram um modelo de análise de sentimentos ao nível da frase capaz de lidar com frases com mais do que uma opinião em que, a cada palavra de sentimento positiva é atribuído um valor +1 e para as negativas -1. O sentimento geral presente na frase é obtido pela soma dos vários valores.

Na maioria das vezes, a implementação deste nível de análise é efetuada em dois problemas de classificação distintos. De referir:

1. **Classificação Subjetiva** – Classifica cada frase como objetiva ou subjetiva. Qualquer método de classificação subjetiva anteriormente apresentado pode ser utilizado.

2. **Classificação da Frase** – Determina a polaridade de cada frase subjetiva. Os métodos supervisionados, não-supervisionados e dicionários, anteriormente apresentados, podem ser aplicados. Para frases com mais do que uma opinião pode ser aplicado o mesmo método utilizado por *Hu e Liu* atrás descrito. *Kim e Hovy* [17] utilizaram um método muito similar mas o sentimento geral das frases é obtido pela multiplicação dos valores de cada sentimento de cada frase.

A utilização deste nível de classificação não é muito diferente do nível do documento. Como referido anteriormente, um documento de análise pode ser desde um documento longo até uma simples frase. Tal como anteriormente, neste último caso, não existe diferença entre nível do documento e nível da frase. O nível da frase também pode ser utilizado para classificar a opinião geral de um documento, utilizando, por exemplo, a soma dos valores da polaridade de cada frase do documento (idêntico ao método de *Hu e Liu* mas para as frases). A semelhança que o nível da frase possui em relação ao nível do documento faz com que este nível seja muitas vezes considerado um passo intermédio no processo de análise de sentimentos se também forem consideradas algumas insuficiências comuns a este nível como:

- Uma frase pode exprimir um tom geral positivo ou negativo mas alguns aspetos da frase podem ter sentidos opostos como no exemplo 2 abaixo que tenta passar um sentimento positivo, contudo, o aspeto "desemprego" é negativo e o aspeto "economia" é positivo.

  Exemplo 2: *"Apesar do grande nível de desemprego, a economia está a ser boa"*

- Tal como no nível do documento e tendo em conta o ponto acima, muitas vezes este nível de classificação não é suficientemente específico para algumas aplicações, nas quais seja necessário saber quais os aspetos positivos e negativos de cada entidade.

2.6.3 Nível da Entidade e do Aspeto

Este nível de classificação vai de encontro à definição de opinião anteriormente apresentada. O alvo de análise agora é muito mais específico do que nos níveis anteriores, focando-se nas entidades e nos seus aspetos. Isto permite analisar qualquer documento mesmo que contenha mais que uma opinião, e com polaridades distintas, e de várias entidades, permitindo ainda analisar com mais detalhe os alvos das opiniões em vez de classificar um documento inteiro baseado na maioria dos seus sentimentos.

Os documentos de análise são divididos em frases como no nível da frase e cada frase é dividida segundo as entidades nela presentes e respetivos aspetos. Os exemplos 3 e 4 representam este processo [18].

Exemplo 3: *"A qualidade do ecrã deste telemóvel é muito boa"* – Entidade: "deste telemóvel"; Aspeto: "ecrã"

Exemplo 4: *"As câmaras Canon têm uma qualidade de imagem ótima"* – Entidade: "câmaras Canon"; Aspeto: "qualidade de imagem"

Este nível de classificação é efetuado em duas fases: extração dos aspetos e classificação dos aspetos.

Extração dos Aspetos

Neste processo identificam-se as entidades e respetivos aspetos presentes em cada frase. Os aspetos são frequentemente representados por substantivos e frases nominais. A extração da entidade é muitas vezes ignorada caso esta seja conhecida à priori, o que implica que apenas uma entidade seja descrita no documento.

Tal como as opiniões, existem aspetos explícitos e implícitos. Estes últimos são mais difíceis de resolver e são expressos por palavras de sentimento que têm esta capacidade dual [18]. Por exemplo, a palavra "silencioso" do exemplo 2 da secção 2.5.1 é uma palavra de sentimento mas também implica o aspeto "som/ruído". A extração de aspetos explícitos pode ser efetuada utilizando algumas estratégias como:

- **Extração baseada em substantivos frequentes e frases nominais** – aplicado em [14] onde é utilizado partes do discurso para identificar este tipo de palavras e apenas as mais frequentes são utilizadas para representar os aspetos; neste processo são utilizadas as palavras de maior frequência, dado que são as que representam os aspetos verdadeiramente importantes que são os que mais são referidos.

- **Extração de substantivos e frases nominais baseada na relação entre opiniões e aspetos** – também utilizada em [14], este tipo de extração assenta no facto de que uma palavra de sentimento é referente a um aspeto ou uma entidade de análise como um todo; dada esta relação, a presença de uma palavra de sentimento significa que na sua proximidade existe um aspeto, logo, é extraído o substantivo ou frase nominal que esteja mais próximo da palavra de sentimento.

- **Extração baseada em modelos supervisionados** – a extração de aspetos pode ser efetuada como um processo supervisionado de extração de informação que naturalmente necessita de um conjunto de dados anotados para treino; o método mais frequentemente utilizado é baseado em aprendizagem sequencial (*sequential learning*) pela utilização de *Hidden Markov Models (HMM)* e *Conditional Random Fields (CRF)*.

Classificação dos Aspetos

Este processo classifica os sentimentos de cada aspeto como positivos, negativos ou neutros. A classificação dos sentimentos pode ser efetuada pela utilização de qualquer um dos modelos já apresentados. A título de exemplo, é apresentado um possível modelo não-supervisionado (maioritariamente baseado em léxicos) utilizado em [11] e constituído por 4 etapas principais [18]. Considere-se a seguinte frase como exemplo de aplicação: *"A qualidade de voz deste telemóvel não é boa mas a bateria é longa"*. Para processarmos esta frase, precisamos de realizar as seguintes operações:

1. **Identificar palavras e frases de sentimento** – para cada frase subjetiva, são identificadas as palavras e frases de sentimento, sendo associada uma pontuação de +1 para as positivas e -1 para as negativas.

   A frase de exemplo ficaria assim: *"A qualidade de voz deste telemóvel não é* ***boa[+1]*** *mas a bateria é longa"*.

   Note-se que a palavra "longa", por si só, não é uma palavra de sentimento mas, quando aplicada num contexto, o sentimento pode ser inferido (etapa 3). "Longa" é, portanto, uma palavra de sentimento dependente do contexto.

2. **Aplicar modificadores de sentimento** – a aplicação destes modificadores altera a pontuação das palavras de sentimento que sejam afetadas - estes modificadores já foram apresentados anteriormente.

   A frase de exemplo ficaria agora: *"A qualidade de voz deste telemóvel não é* ***boa[-1]*** *mas a bateria é longa"*.

3. **Tratar palavras de contraste (*but-clauses*)** – palavras ou frases que impliquem contrariedade necessitam de ser tratadas de forma especial porque, normalmente, alteram a polaridade dos sentimentos. As palavras de contraste mais comuns são "mas" e "contudo" e as frases são "com exceção de", "exceto aquilo", "exceto por".

   A utilização deste método está sujeita à regra: a polaridade dos sentimentos, imediatamente antes e depois da palavra de contraste, é oposta se não for possível determinar a polaridade de um dos lados.

   A frase de exemplo ficaria: *"A qualidade de voz deste telemóvel não é **boa[-1]** mas a bateria é **longa[+1]**"*.

   Existem casos onde a presença de palavras de contraste não implica a existência de sentimentos contrários como em *"não só ... **mas** também"*. Estes casos têm de ser identificados previamente.

4. **Agregar as pontuações** – As pontuações obtidas nas palavras de sentimento são agregadas de acordo com um certo método ou fórmula. Vários autores utilizam diferentes formas de agregação, desde a utilização de fórmulas matemáticas baseadas na distância entre palavras, como no caso do autor original do modelo, até à simples soma das pontuações, como em [14].

   Caso a pontuação final seja positiva então o documento de análise é classificado como positivo, caso seja negativa é classificado como negativo e é neutro caso contrário. Para a frase de exemplo, o sentimento geral seria neutro.

2.7 Problemas e Dificuldades Gerais

A informação até agora exposta é referente, de uma forma geral, à definição e a vários processos envolvidos em processo de análise de sentimentos. Como tal, é independente de qualquer aplicação sobre um problema ou contexto específico. Em conjunto com a informação foram apresentados vários exemplos, alguns representativos de casos específicos de análise e de dificuldades que representam alguns dos problemas gerais da análise de sentimentos.

Vários autores definiram um conjunto de problemas gerais a qualquer solução de análise de sentimentos, independentemente do problema que se procura tratar. Note-se que a aplicação num contexto de análise específico pode originar problemas inerentes a esse contexto. A compreensão das limitações e problemas gerais de uma área como análise de sentimentos é de extrema importância para o desenvolvimento de qualquer solução da mesma. De seguida

são apresentados alguns desses problemas. A ordem com que se apresentam esses problemas não reflete qualquer nível crescente ou decrescente de dificuldade. Os problemas são, então, os seguintes:

1. **Frases Objetivas** – Nestas frases as opiniões são expressas em conteúdo subjetivo com a utilização de palavras ou frases características do mesmo. Contudo, o conteúdo subjetivo não é o único onde sentimentos podem ser expressos. As frases objetivas, características por apresentar maioritariamente factos, também podem exprimir sentimentos como foi visto no exemplo 1 da secção 2.4. Isto deve-se à possível presença de factos desejáveis ou indesejáveis expostos em factos. Abaixo apresenta-se um outro exemplo deste problema.

   Exemplo 1: *"Comprei o colchão ontem e hoje tem uma curvatura a meio"*

2. **Contextos** – Este é um dos problemas essenciais referente ao contexto de aplicação de uma solução de análise de sentimentos. Uma solução de análise de sentimentos é muito dependente do contexto para o qual é desenvolvida. Isto leva a uma dificuldade de utilização fora do contexto original. Este problema tem a sua origem na construção de uma solução na qual são utilizados métodos léxicos e de aprendizagem automática que fazem uso de recursos anotados. Estes recursos são construídos com informação do contexto específico onde a solução vai ser aplicada. A aplicação da solução noutros contextos pode levar a uma perda de precisão, uma vez que algumas palavras podem ter polaridades opostas em contextos diferentes como nos exemplos 1 e 2 da secção 2.5 anteriormente apresentados.

3. **Sarcasmo** – O sarcasmo é uma forma de exprimir opiniões que, até para as pessoas, pode ser de difícil distinção, dado que é característico por apresentar o oposto ao que de facto se pretende. Em análise de sentimentos isto significa que quando é expresso um sentimento positivo na realidade o autor refere-se a um negativo e vice-versa. A deteção de sarcasmo por parte de um modelo computacional é, portanto, problemático. O exemplo 6 da secção 2.5 é representativo de sarcasmo. Existem alguns trabalhos realizados no sentido de resolução de informação sarcástica como em [12] e [10].

4. **Negações** – A correta resolução de negações é um problema de grande dificuldade e importância dado que negações frequentemente alteram a polaridade dos sentimentos. A não deteção de uma negação pode, assim, alterar completamente o resultado da polaridade de um documento de análise ou de partes do mesmo. As negações são frequentemente representadas por palavras de negação como "não" ou frases específicas de contexto. Porém não se pode resumir à pura deteção desse tipo de palavras no documento de análise porque estas palavras nem sempre alteram a polaridade dos sentimentos como no exemplo abaixo onde a presença da palavra "só" depois da palavra de

negação faz com que a mesma não altere a polaridade presente na frase.

Exemplo 2: *"Não só ... mas também"*

5. **Comparações e Frases Interrogativas e Condicionais** – A utilização de comparações é uma outra forma de expressar sentimentos pela comparação de aspetos de duas ou mais entidades. A análise de sentimentos em frases comparativas requer uma abordagem diferente das frases regulares.

   Exemplo 3: *"Mercedes tem melhores motores do que BMW"*

   O exemplo acima não indica que a qualidade de motor de uma ou outra marca é boa ou má mas limita-se a compará-los. No entanto, existe um sentimento positivo em relação à primeira marca referente à preferência do autor. Por outro lado, e tal como em frases regulares, uma frase comparativa nem sempre possui sentimentos, como no exemplo abaixo, onde é apenas apresentado factos das entidades.

   Exemplo 4: *"Mercedes são maiores do que BMWs"*

   Para além de toda a dificuldade de extração das entidades e dos aspetos de cada uma e ainda da classificação dos sentimentos, o primeiro problema das frases comparativas é a correta identificação das mesmas com distinção das que de facto apresentam sentimentos devido à variedade de formas como uma comparação pode ser efetuada.

   Para além de frases comparativas, frases interrogativas e condicionais apresentam alguns problemas dado que podem conter palavras de sentimento mas não exprimir qualquer sentimento como nos exemplos abaixo.

   Exemplo 5: *"Qual a melhor câmara da Sony?"*

   Exemplo 6: *"Se encontrar uma boa câmara na loja, vou compra-la"*

   Como seria de esperar, algumas frases deste tipo exprimem sentimentos como no exemplo abaixo. Embora a identificação deste tipo de frases seja na maioria das vezes fácil, a identificação das que de facto expressam sentimentos mostra ser uma tarefa difícil.

   Exemplo 7: *"Se queres um bom carro, compra um Rolls Royce"*

6. **Linguagem** – Os modelos que são desenvolvidos para uma dada linguagem não podem ser aplicados sobre outras linguagens. A maior parte dos recursos, como dados de treino e estudos linguísticos e de escrita, foram desenvolvidos para a língua inglesa, o que torna difícil a aplicação noutras linguagens. Considere-se, por exemplo, uma empresa que possui uma solução de análise de sentimentos direcionada para o inglês, mas que,

também, necessita da mesma solução para as línguas dos restantes mercados onde se encontra.

Vários autores têm estudado este problema como em [35], mas o mesmo está longe de ser resolvido ou de ser encontrada uma solução, na qual a perda de precisão da solução na língua original para a nova seja mínima. A maioria das soluções passa por recorrer à tradução de conteúdo no qual podem ser colocadas duas hipóteses principais. A primeira envolve a tradução do conteúdo de análise da língua pretendida para a língua utilizada no modelo existente. A segunda é a tradução do conjunto de dados utilizados para construção do modelo para a língua pretendida e só depois construir o modelo obtendo-se assim um modelo específico para a linguagem pretendida. Apesar destas soluções serem as mais utilizadas é fácil compreender que qualquer uma delas leva a uma perda de precisão nos resultados devido a erros no processo de tradução e a várias nuances entre as duas linguagens envolvidas.

7. **Pronomes** – A resolução de pronomes é um processo difícil para qualquer sistema computacional. As opiniões e sentimentos refletidos sobre algo que seja representado por um pronome torna difícil identificar o alvo desse pronome.

   Exemplo 8: *"Fomos ao cinema e depois fomos jantar. Não gostei nada dele"*

   O exemplo acima é característico deste problema onde o pronome "dele" é utilizado para expressar um sentimento sobre um alvo que neste caso é o cinema ou o jantar. Sem qualquer outra informação, a resolução do pronome do exemplo apresentado é impossível mesmo para uma pessoa.

8. **Avaliação** – Apesar de este não ser um problema relacionado com o desenvolvimento de um modelo de análise de sentimentos, é um problema comum a qualquer modelo relacionado com a forma de avaliação da precisão obtida. Dado que cada modelo é desenvolvido para um contexto específico, a avaliação do mesmo normalmente passa pela análise da precisão obtida pelo modelo sobre conjuntos de dados de treino do contexto em que se insere. Isto é bastante problemático porque, embora seja possível testar o modelo em relação à sua performance, não existe uma forma universal de comparar os vários modelos desenvolvidos, porque cada um deles está inserido, potencialmente, em contextos de análise distintos. Note-se ainda que modelos desenvolvidos num mesmo contexto, necessitam de ser desenvolvidos e testados sobre o mesmo conjunto de dados para que a precisão de cada um dos modelos possa ser comparada de forma precisa.

3 O Pré-Processamento de Dados

3.1 Introdução

Grandes quantidades de dados são gerados diariamente por uma grande diversidade de fontes. À medida que esta quantidade de dados vai aumentando, aumenta também a possibilidade de diminuição da qualidade dos dados. As fontes devem de alguma forma garantir, tanto quanto possível, a boa qualidade dos dados que por elas são gerados. A qualidade dos dados está dependente da capacidade e forma como as fontes são desenvolvidas. Uma fonte desenvolvida com o objetivo de dar suporte, para além das funcionalidades básicas que representa, como a geração de dados para um sistema de *data warehousing (DW)*, frequentemente, tem a capacidade de produzir um conjunto de dados bem estruturados e de qualidade superior a outras fontes desenvolvidas para sistemas menos cruciais. Isto não invalida, contudo, a capacidade de geração de dados de boa qualidade por parte de pequenas fontes. A qualidade de um conjunto de dados é, portanto, dependente da fonte responsável pela geração dos mesmos. Porém, independentemente da fonte, existe a possibilidade de ocorrência de inconsistências nos dados. Estas inconsistências devem ser detetadas e tratadas na fase de pré-processamento dos dados.

O crescimento exponencial que as tecnologias têm sofrido proporciona mesmo aos pequenos sistemas a utilização de grandes conjunto de dados. Muitos são os casos nos quais os dados utilizados por um qualquer sistema não estão preparados ou adequados em termos de presença de inconsistências e de formato. A fase de pré-processamento dos dados possui um papel fundamental para garantir que tal não aconteça. O tratamento de inconsistências e de formatação dos dados, são muito comuns no pré-processamento de dados. A importância desta fase aumenta quando é necessário também fazer a transformação do conjunto de dados, inserindo ou removendo novas variáveis, através de ações de análise e de compreensão dos dados, que também fazem parte integrante do pré-processamento, e que permitem obter uma visão global sobre os dados pela análise da forma como as variáveis se relacionam entre si e sendo responsável por qualquer decisão posterior de transformação aos dados.

A figura 3.1 representa o típico esquema de pré-processamento de dados aplicado no desenvolvimento da maioria dos sistemas. O esquema apresentado é, na verdade, uma simplificação

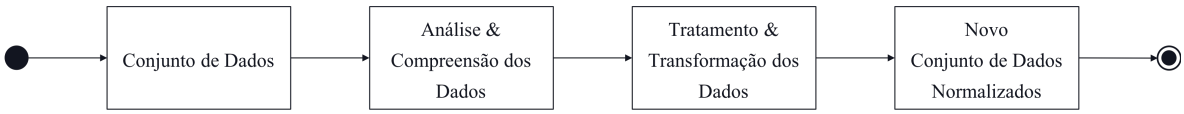


Figura 3.1: Pré-processamento típico.

representativa da metodologia padrão aplicada globalmente em soluções de análise de dados, que foi desenvolvida pela IBM – o CRISP-DM [15]. Não diferente ao que seria de esperar, a abordagem utilizada para o desenvolvimento das soluções propostas no presente capítulo e posteriores segue os padrões básicos desta metodologia.

A importância do pré-processamento vai ainda mais além do que aquilo que já foi exposto. O correto planeamento de todas as fases constituintes permite garantir um controlo e compreensão total do conjunto de dados, prever o comportamento dos dados no futuro, facilitar o desenvolvimento de modelos de aprendizagem automática, entre outros, e especialmente, garantir que não é necessário voltar a esta fase depois da construção de modelos ou de sistemas cujo comportamento e resultados não estejam de acordo com o esperado, o que pode ser justificável por uma deficiência no pré-processamento dos dados utilizados. Esta fase representa, assim, a maioria do tempo gasto no desenvolvimento de uma solução de análise de dados, mesmo quando comparado com a implementação de modelos de aprendizagem automática que atuam sobre os dados.

3.2 Dataset

Um Dataset é um conjunto de dados que é utilizado para modular um sistema. Formalmente, um conjunto de dados pode ser definido como uma tabela de dados, na qual cada coluna representa uma variável e cada linha representa uma entrada de informação sobre essa tabela. Pode ser ainda definido de forma mais simples como o processo de recolha de informação (texto, imagem, som) e armazenamento da mesma de uma forma mais ou menos estruturada. Embora aceitáveis, estas definições, de uma forma generalizada, tornam difícil definir globalmente qualquer conjunto de dados, devido ao tipo e à forma como os dados são armazenados e à própria natureza da fonte (ex.: bases de dados relacionais vs bases de dados não-relacionais).

A escolha do conjunto de dados a utilizar num processo de modulação deve representar, tanto quanto possível, o objetivo do sistema final, quer em termos da qualidade como da quantidade de dados. Para além disto, um conjunto de dados com uma boa diversidade de casos possíveis do contexto é fundamental para a compreensão do problema, quer em termos dos modelos de aprendizagem automática como em termos de soluções manuais baseadas em algoritmos. A

quantidade e a diversidade dos dados são duas características que devem estar presentes de forma equilibrada – uma elevada quantidade de dados e uma baixa diversidade de dados origina um desbalanceamento das classes do dataset e dos problemas associados com ele, enquanto que uma baixa quantidade de dados com alta diversidade (tanto quanto possível dada a baixa quantidade) pode levar a que o dataset não seja suficientemente grande para o treino e teste de modelos de aprendizagem automática. Como referido na secção anterior, a escolha de uma boa fonte permite garantir que os dados tenham este equilíbrio necessário entre qualidade, quantidade e diversidade.

Dada a grandeza da Internet e a quantidade de projetos de investigação desenvolvidos no domínio da análise de dados, existem muitas fontes que disponibilizam datasets preparados, com qualidade conhecida, sobre os mais diversos temas e contextos de análise. Chegam a existir mesmo datasets universais utilizados para avaliação de modelos de acordo com um dado contexto. Isto é especialmente visível para datasets em língua inglesa, mas muito menos visível para outras línguas – algo que levantou algumas dificuldades no desenvolvimento do sistema desenvolvido[1]. Como tal, a procura de um dataset apropriado para o contexto de análise no presente projeto não resultou em nada resultados satisfatórios, devido à necessidade da língua portuguesa. Em contrapartida, um dataset em língua inglesa era facilmente obtido. De facto, isto é um problema muito comum e que já foi exposto anteriormente no ponto 6 da secção 2.7. Assim, foram consideradas duas possibilidades para obter o dataset pretendido:

1. Tradução automática para português de um dataset em inglês.

2. Construção de um dataset de raíz.

A primeira possibilidade é uma das mais utilizadas para contornar este tipo de problema. Porém, possui a desvantagem da inevitável ocorrência de erros no processo de tradução que podem diminuir posteriormente a precisão dos modelos de aprendizagem automática. Por este motivo, foi escolhida a segunda possibilidade.

3.2.1 A Construção de um Dataset

Dada a natureza do tipo de dados a serem utilizados é possível encontrar facilmente na Internet um grande conjunto de comentários provenientes de várias fontes. Neste sentido, foi efetuada uma análise às possíveis fontes e dados, que encontrámos disponíveis na Internet, para escolha da qual melhor representasse o contexto de análise e que possuísse a melhor relação entre

[1] Análise de sentimentos em conteúdos textuais, especificamente comentários de produtos em língua portuguesa. Os produtos pertencem ao contexto de eletrodomésticos e acessórios e eletrónicos (ex.: computadores, smartphones, tv's, etc) e acessórios.

qualidade, quantidade e diversidade dos dados. A fonte escolhida, cujo nome não é revelado por questões de privacidade e porque não acrescenta qualquer relevância, é um dos principais websites de venda de produtos. Nesse website encontrámos os contextos desejados para análise. A tabela 3.1 revela os contextos gerais dos produtos que encontrámos no site referido.

Contextos dos Produtos
Eletrodomésticos
Beleza, Saúde e Fitness
Tv, Vídeo e Som
Fotografia
Informática
Acessórios de Informática
Smartphones e Comunicações
Gaming

Tabela 3.1: Contextos gerais dos produtos do dataset.

O website utilizado não possui qualquer *API* para facilitar o processo de recolha de dados. Devido a este facto, foi necessário desenvolver um *web scraper* para extrair de forma automática todos os comentários realizados sobre todos os produtos presentes no website. A figura 3.2 apresenta o esquema de funcionamento do web scraper, desenvolvido especificamente para o website selecionado, tendo a capacidade de ser totalmente automático na extração e armazenamento dos dados. De entre a informação disponível no website em questão, apenas se procedeu à extração do seguinte[2]:

1. Número de estrelas (1 a 5) atribuídas pelo autor do comentário ao produto alvo.
2. Data em que o comentário foi escrito.
3. Comentário realizado.

O elemento de dados "data" foi extraído apenas como uma variável extra de complemento ao dataset para suporte a uma eventual necessidade de análises temporais por parte de um outro sistema (ex.: sistemas analíticos necessitam de uma dimensão temporal). O número de estrelas foi extraído com o objetivo de ser utilizado em conjunto com os respetivos comentários, originando assim um conjunto de dados anotados para treino e teste de modelos supervisionados de aprendizagem automática. Os dados extraídos foram armazenados, de forma estruturada, num ficheiro *.csv* para posterior análise e tratamento.

[2]Não foi extraído qualquer outro tipo de informação como o nome do autor do comentário ou o local onde foi efetuado, não é realizado qualquer tipo de *profiling* aos autores dos comentários nem o sistema desenvolvido é utilizado para qualquer efeito comercial.

Figura 3.2: Web scraper.

3.2.2 Análise Inicial

A realização de uma análise inicial básica sobre o conjunto de dados é fundamental para perceber o seu tamanho, o número de atributos envolvidos e o seu tipo, a qualidade dos dados, entre outros. O conjunto de dados possui 12590 comentários. A tabela 3.2 mostra as primeiras linhas do dataset. O nome dos atributos é auto explicativo, com a possível exceção do primeiro deles, que representa o número de estrelas atribuídas aos comentários.

Rate	Data	Comentário
5	10/01/18	Estou muito satisfeita com esta maquina, é excelente.
1	22/12/17	Tem junta Anti-bacteriana que evita manchas e bolores mas a borracha do oculo da porta são muito fragéis, borracha solta e perda de água passado 7 meses de adquerir, a espera de assistência.
5	22/12/17	Poderia ser de mais fácil utilização os programas nem sempre são fáceis de utilizar
4	14/12/17	Relativamente ao escoamento da água na borracha da porta não é funcional pois acumula água. A porta de abertura poderia ter mais ângulo de abertura.
4	06/08/17	Estou bastante contente com esta máquina de lavar roupa

Tabela 3.2: Extrato das primeiras linhas do dataset.

Linhas	12590
Colunas	3

Tabela 3.3: Dimensões do dataset.

O principal objetivo de um sistema de análise de sentimentos, independentemente do contexto para o qual é desenvolvido, é a previsão da polaridade de um dado documento. A polaridade pode ser representada numa escala nominal, como é o caso do número de estrelas de cada comentário, ou numa escala binária. No primeiro caso é necessário uma correspondência entre os valores da escala que são considerados positivos e negativos para que possua um significado. Em contrapartida, numa escala binária, o significado é imediato. Porém, não possui o mesmo nível de informação presente numa escala nominal. A escolha da escala a utilizar depende, normalmente, do tipo de modelos que se pretende implementar e do nível de classificação (documento, frase, entidade ou aspeto).

A maioria dos sistemas de análise de sentimentos utiliza uma escala binária para poderem utilizar modelos supervisionados de classificação de aprendizagem automática. Não diferente, as soluções de aprendizagem automática desenvolvidas (e apresentadas posteriormente) utilizam também esta técnica de modelação, o que fez com que se tivesse de anotar a polaridade dos comentários numa escala binária. Para tal, foi efetuada uma transformação do número de estrelas de cada comentário para uma escala binária, de acordo com a correspondência apresentada na tabela 3.4. O resultado da transformação é acrescentado ao dataset num novo atributo – "Polaridade" – com uma codificação realizada de acordo com a tabela 3.5. A tabela 3.6 mostra as primeiras linhas do dataset, já com o novo atributo (a data foi omitida).

Nº Estrelas	Polaridade
1	Negativo
2	Negativo
3	Análise manual
4	Positivo
5	Positivo

Tabela 3.4: Transformação binária do número de estrelas.

Positivo	0
Negativo	1

Tabela 3.5: Polaridade binária dos comentários.

Rate	Polaridade	Comentário
5	0	Estou muito satisfeita com esta maquina, é excelente.
1	1	Tem junta Anti-bacteriana que evita manchas e bolores mas a borracha do oculo da porta são muito fragéis, borracha solta e perda de água passado 7 meses de adquerir, a espera de assistência.
5	0	Poderia ser de mais fácil utilização os programas nem sempre são fáceis de utilizar
4	0	Relativamente ao escoamento da água na borracha da porta não é funcional pois acumula água. A porta de abertura poderia ter mais ângulo de abertura.
4	0	Estou bastante contente com esta máquina de lavar roupa

Tabela 3.6: Extrato das primeiras linhas do dataset com polaridade binária.

A transformação da tabela 3.4 é relativamente óbvia. Numa escala de 1 a 5, onde a positividade aumenta com o número de estrelas, as estrelas 1 e 2 podem ser automaticamente consideradas negativas, bem como as estrelas 4 e 5 podem ser consideradas positivas. A maior dificuldade ocorre na transformação da estrela 3, que por ser intermédia, torna difícil obter uma correta

transformação automática. Neste caso, foi efetuada uma análise manual a cada comentário com três estrelas para determinar a sua polaridade, de forma o mais precisa possível, dada a grande subjetividade que os sentimentos costumam revelar.

Agora que o conjunto de dados está anotado em classes binárias – Positivo (0) e Negativo (1) para cada comentário – é possível obter uma sensibilidade da distribuição dos comentários pelas classes. A figura 3.3 mostra que existem 10828 comentários positivos (86% de todos os comentários) e 1762 comentários negativos (14% de todos os comentários).

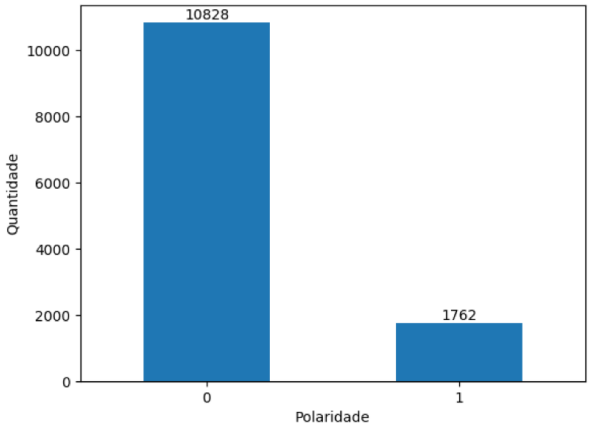


Figura 3.3: Distribuição dos comentários por polaridade.

Os valores da distribuição das classes mostram o primeiro possível problema relacionado com a qualidade dos dados: o não balanceamento das classes. Isto é especialmente importante na modelação de soluções de aprendizagem automática, nas quais o não balanceamento do atributo (ou atributos) utilizados no processo de aprendizagem pode influenciar o modelo, tornando-o tendencioso para a classe maioritária. No processo de modelação das soluções de aprendizagem automática é necessário verificar se tal acontece, analisando para além do valor da precisão, que embora possa ser elevado, os valores de outras métricas como sensibilidade, especificidade, ou a área de baixo da curva ROC (AUROC). Estes elementos podem revelar a presença deste problema e, como tal, devem ser utilizados como forma de avaliação do modelo. Informação adicional acerca do não balanceamento das classes será apresentada mais tarde no capítulo de desenvolvimento dos modelos de aprendizagem automática.

Ainda em relação à qualidade dos dados, não foi detetado qualquer valor omisso, o que indica o correto funcionamento do web scraper desenvolvido. Dado que, de entre os atributos do dataset apenas dois desses atributos são relevantes para utilização ("Polaridade" e "Comentário"), e uma vez que não existem valores omissos e que o atributo "Polaridade" é binário, toda a restante análise e tratamento dos dados vai estar relacionada com a qualidade dos comentários – ver próxima secção.

3.3 Pré-Processamento

Após uma análise inicial ao conjunto de dados é necessário uma análise mais detalhada para que se possa garantir uma qualidade de dados uniforme, sobretudo sobre os comentários que foram realizados. Para além de assegurar a qualidade dos dados envolvidos, esta análise vai permitir também uma melhor compreensão da forma como os sentimentos são expressos em comentários e ainda a implementação de um conjunto de transformações que representam a resolução dos vários problemas de qualidade, bem como possíveis novas formas de direcionar o processo de análise. Toda a análise aqui efetuada vai servir de base para decidir quais as transformações a aplicar ao conjunto de dados em cada uma das solução implementadas (aprendizagem automática, baseada em dicionários de sentimentos e híbrida).

3.3.1 Novos Atributos

Dado que existem apenas dois atributos no dataset que são relevantes para análise, numa primeira instância tentou-se adicionar novos atributos para aumentar o potencial de aprendizagem, sobretudo para os modelos de aprendizagem automática supervisionados. O motivo para tal está relacionado com a forma de funcionamento dos modelos de aprendizagem automática, nos quais a existência de mais atributos (até um certo ponto) que expliquem o atributo de análise, neste caso a polaridade dos comentários, ajuda a convergir mais eficazmente o modelo. Neste sentido, foi criado um novo atributo – *ComprimentoComentario* – com o comprimento de cada comentário para tentar perceber se existe alguma relação significativa entre o comprimento dos comentários e a polaridade.

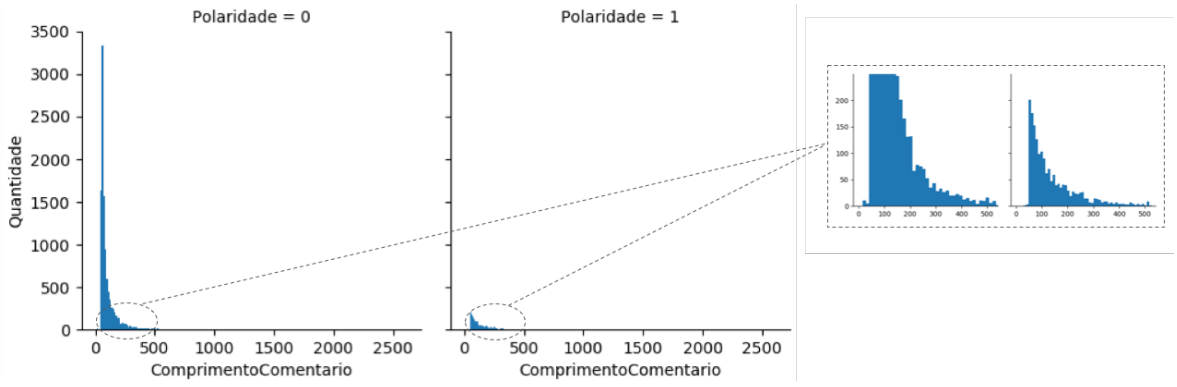


Figura 3.4: Comprimento dos comentários por polaridade (histograma).

As figuras 3.4 e 3.5 mostram a distribuição do comprimento dos comentários por polaridade. É possível observar que a distribuição do comprimento dos comentários segue um padrão pra-

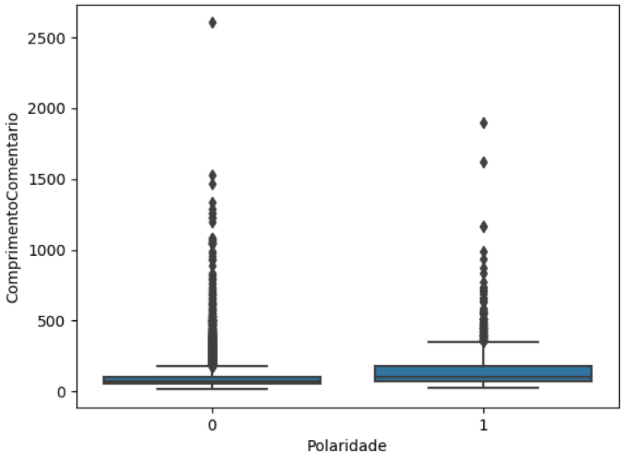


Figura 3.5: Comprimento dos comentários por polaridade (boxplot).

ticamente idêntico para qualquer polaridade (figura 3.4) e a mediana dos comprimentos pelas polaridades é, também, muito próxima (figura 3.5). Note-se ainda que na figura 3.4 os valores da quantidade são muito mais elevados para as polaridades positivas do que para as polaridades negativas, devido ao não balanceamento dos dados, anteriormente discutido, no qual existe um maior número de comentários positivos do que negativos. Estas observações permitem sustentar que não existe uma relação significativa entre o comprimento dos comentários e a sua polaridade. Tal facto é confirmado pela matriz de correlação dos atributos (numéricos), na figura 3.6, na qual os valores da correlação entre o comprimento dos comentários e a polaridade são muito baixos. Dado estes factos, a utilização do comprimento dos comentários foi descartada da restante análise e de qualquer solução implementada.

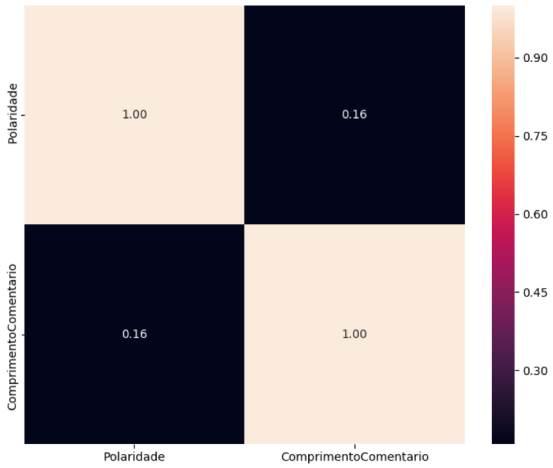


Figura 3.6: Matriz de correlação – comprimento dos comentários/polaridade.

3.3.2 Análise, Transformação e Qualidade dos Comentários

Os comentários são o atributo essencial de análise. É através deles que os sentimentos são expressos, tornando fundamental garantir um formato e qualidade universal entre si. Neste sentido, a parte da análise que nos falta abordar vai incidir sobre os comentários e segue, de uma forma geral, a estrutura apresentada na figura 3.7.

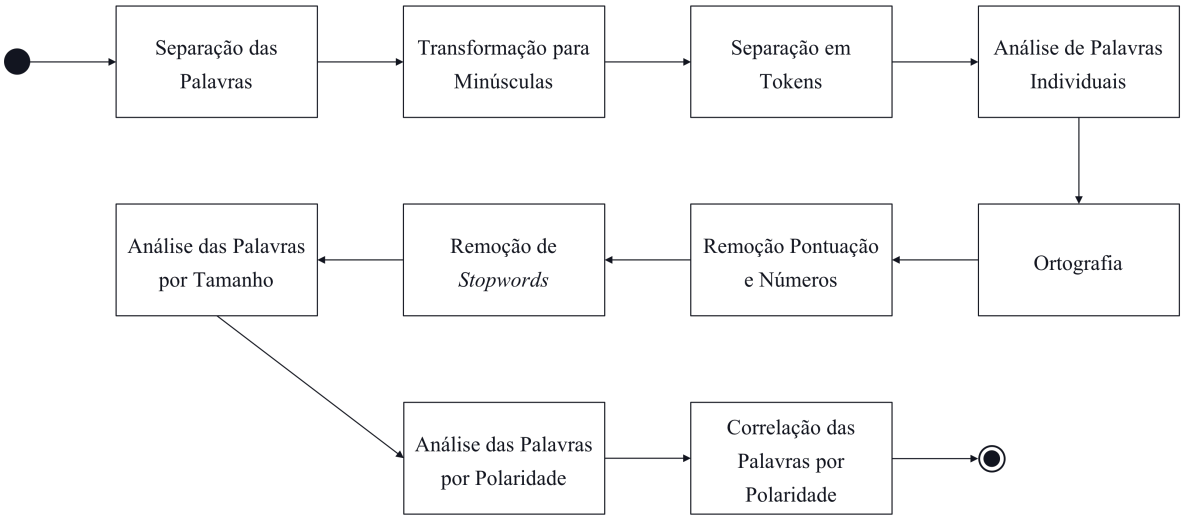


Figura 3.7: Estrutura do pré-processamento.

Separação das Palavras

A qualidade da escrita dos comentários é o maior problema que enfrentamos na sua análise. Dada a origem do conjunto de dados, o formato de escrita é caracteristicamente descuidado, reduzindo significativamente a qualidade dos próprios comentários. Este facto vai para além deste ponto de análise, estando presente em todo o pré-processamento dos dados. Considere-se os seguintes comentários presentes no conjunto de dados:

1. *Boas funcionalidades.Muito eficaz nas lavagens de curta duração, e um modelo espetacular! 5******

2. *Bom aparelho,com bom som,e facil de ligar,com comando facil de utilizar.*

3. *AdoreiPara mim foi a melhor compra que fiz ate hoje*

A observação dos comentários acima apresentados revela vários problemas de qualidade. Para o presente ponto de análise interessa apenas os problemas relacionados com a incorreta junção de duas palavras distintas, como é possível ver a sublinhado nos comentários. Embora seja

possível compreender o significado destes casos, para um computador/sistema estes casos são vistos como uma única palavra alterando o significado e estrutura do comentário. Estes casos foram resolvidos pela sua deteção automática e adição de um espaço entre qualquer sinal de pontuação e uma letra minúscula ou maiúscula e entre qualquer letra minúscula seguida por uma letra maiúscula.

Transformação para Minúsculas e Tokens

Neste ponto foram implementadas duas transformações simples mas essenciais para a análise:

1. Transformação dos comentários para palavras minúsculas
2. Representação interna dos comentários em tokens

A transformação para palavras minúsculas é uma forma de uniformizar cada vez mais cada comentário. Esta transformação permite que uma mesma palavra não seja considerada como duas palavras distintas (ex.: Gostei vs gostei). Consegue-se, assim, diminuir o vocabulário[3] de todo o conjunto de dados, uma vez que palavras anteriormente consideradas distintas passam a ser a mesma palavra.

Na figura 3.8 podemos ver o tamanho do vocabulário antes e depois de aplicar esta transformação. A relevância desta transformação tem um impacto fundamental tanto para as soluções baseadas em aprendizagem automática supervisionada como para as baseadas em dicionários de sentimentos. O primeiro caso está relacionado com a forma como os modelos supervisionados utilizam os comentários para aprendizagem, nomeadamente com o uso de vetores de tamanho igual ao vocabulário, em que a redução do vocabulário produz vetores de menor dimensão e redundância aumentando a capacidade de aprendizagem e convergência do modelo. A título explicativo, considere-se o exemplo 1, relacionado com a forma como um modelo supervisionado baseado em *bag of words* utiliza a representação de dois comentários semelhantes sem a transformação para minúscula e com esta mesma transformação. Note-se que o exemplo abaixo representa um caso simples, construído para mostrar a importância da uniformização dos comentários e redução do vocabulário.

[3]Vocabulário é o conjunto de todas as palavras distintas de todos os comentários.

Exemplo 1: Considere-se dois comentários semelhantes C1 e C2:

C1: *Gostei muito do produto*

C2: *gostei mesmo muito do produto*

Sem transformação para minúsculas:

	Gostei	muito	do	produto	**gostei**	mesmo
Vetor C1	1	1	1	1	0	0
Vetor C2	0	1	1	1	1	1

Os comentários são representados em dois vetores distintos.

Com transformação para minúsculas:

	gostei	muito	do	produto	mesmo
Vetor C1	1	1	1	1	0
Vetor C2	1	1	1	1	1

Os comentários são representados em dois vetores distintos mas o tamanho dos mesmos foi reduzido. Uniformização/redução de redundância da palavra "gostei".

Para os modelos baseados em dicionários esta transformação é também importante, dado que permite que o dicionário ou dicionários e restantes recursos utilizados sejam desenvolvidos apenas e garantidamente para um vocabulário em minúsculas.

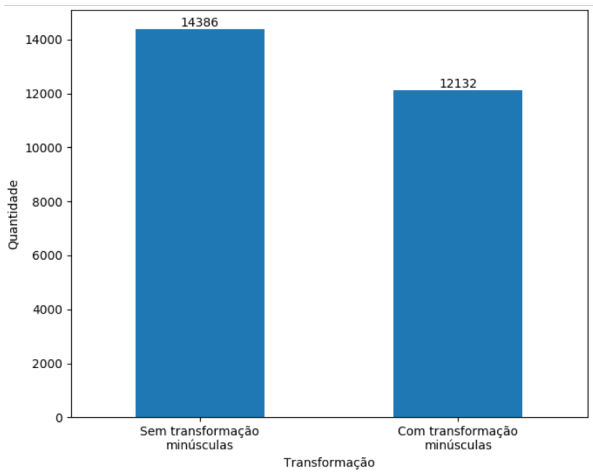


Figura 3.8: Tamanho do vocabulário antes e depois da transformação para minúsculas.

A transformação para *tokens* permite que cada comentário deixe de ser representado como uma *string* passando a ser representado por uma lista de palavras individuais. Esta transformação

facilita a análise individual de cada palavra do vocabulário. O método mais comum para transformação de uma *string* em palavras individuais (*tokens*) é pela separação em espaços. Este método é simples e de fácil implementação. Contudo, não é adequado para o nível de qualidade de análise que se pretende obter, uma vez que qualquer sinal de pontuação mantém-se junto (não é separado) das palavras. Uma outra solução passava por, para além da separação em espaços, separar também em sinais de pontuação. Porém esta solução produz também alguns problemas como quebra de siglas e emojis, que, como sabemos, são representados por sinais de pontuação.

A separação em *tokens* não é, portanto, um processo trivial e o resultado deve de ir de encontro com as necessidades e nível de análise necessários. Neste sentido, foi utilizado o *TweetTokenizer*, da biblioteca NLTK [7], desenvolvido para utilização em *tweets*, mas cujo resultado é aplicável ao nível de análise pretendido. Este tokenizer separa os comentários em palavras individuais, com separação dos sinais de pontuação das palavras e com a capacidade de manter siglas e emojis. O exemplo 2 mostra o resultado da aplicação deste tokenizer a um comentário do dataset.

Exemplo 2: A tabela abaixo mostra a transformação para minúsculas e *tokens* sobre um comentário.

Comentário	*estou satisfeita: qualidade/preço acessível, jarro leve, elegante.... RECOMENDO:-)*
Minúsculas	*estou satisfeita: qualidade/preço acessível, jarro leve, elegante.... recomendo:-)*
Tokenizer	*['estou', 'satisfeita', ':', 'qualidade', '/', 'preço', 'acessível, ',', 'jarro', 'leve', ',', 'elegante', '...', 'recomendo', ':-)']*

Tabela 3.7: Resultado da transformação para letras minúsculas e *tokens*.

Palavras e Ortografia

Até agora, as transformações efetuadas permitem analisar mais facilmente cada palavra de forma individual. Torna-se essencial uma perceção da distribuição das palavras, em particular as que possuem maior frequência, dado que tendem a caracterizar um texto ou um vocabulário. A figura 3.9 mostra a distribuição das 50 palavras mais frequentes dos comentários e a tabela 3.8 mostra estas palavras, pela mesma ordem da distribuição, para facilidade de leitura. A figura e a tabela revelam que, na verdade, nem tudo são palavras. Sinais de pontuação e números (embora não estejam presentes na distribuição acima) são ainda parte constituinte do vocabulário. A análise em detalhe destes elementos é efetuada no próximo ponto sendo o presente para a análise de um fator que influencia de forma significativa a qualidade dos dados: a ortografia.

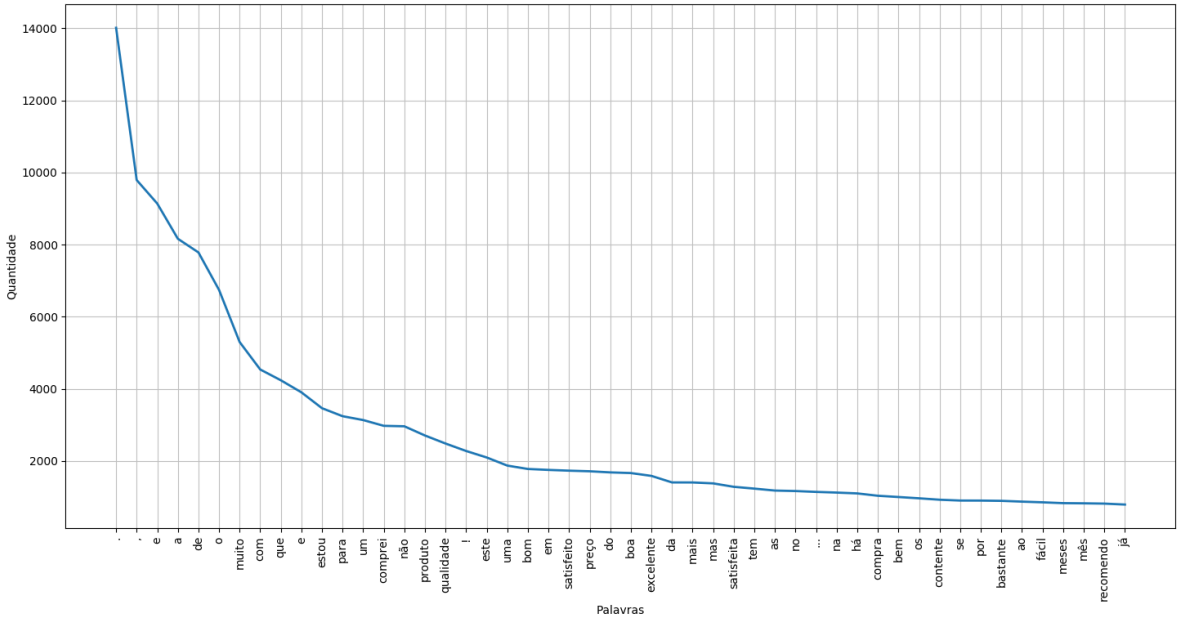


Figura 3.9: Distribuição das 50 palavras mais frequentes dos comentários (com sinais de pontuação e números).

.	,	e	a	de	o	muito	com	que	é
estou	para	um	comprei	não	produto	qualidade	!	este	uma
bom	em	satisfeito	preço	do	boa	excelente	da	mais	mas
satisfeita	tem	as	no	...	na	há	compra	bem	os
contente	se	por	bastante	ao	fácil	meses	mês	recomendo	já

Tabela 3.8: 50 palavras mais frequentes dos comentários (com sinais de pontuação e números).

Dada a natureza descuidada característica do tipo de comentários do conjunto de dados, a presença de erros ortográficos é um fator que ocorre com frequência. Naturalmente, qualquer palavra que contenha um erro ortográfico deixa de possuir o sentido e o papel que desempenha levando à diminuição da qualidade dos dados e dos comentários. Este fator tem um impacto para além da qualidade dos dados, tal como foi visto na transformação para letras minúsculas, uma palavra com um erro ortográfico vai ser interpretada como uma nova palavra (em relação à palavra que na verdade representa) tanto por modelos de aprendizagem automática supervisionada como por soluções baseadas em dicionários de sentimentos levando, no primeiro caso, a um aumento do tamanho dos vetores utilizados para o processo de aprendizagem e à não localização da palavra no dicionário de sentimentos que dependendo da implementação, muito provavelmente leva a uma atribuição de uma polaridade nula sobre a palavra, para soluções em dicionários.

A análise manual das palavras com maior frequência, para além das apresentadas na figura/tabela acima, e ainda das palavras com menor frequência, permitiu obter um conjunto de

palavras com erros ortográficos e uma perceção de possíveis palavras suscetíveis a erros. As palavras com menor frequência tiveram um papel nesta análise dada a hipótese de, apesar da suscetibilidade da ocorrência de erros, os que ocorrem são em pouca frequência em relação à palavra natural[4], apesar da possível grande quantidade de palavras naturais com erros ortográficos. As palavras com maior frequência permitiram que fosse identificado um conjunto de possíveis palavras naturais que devido às suas características como presença de assentos, capacidade de arredondamento (ex.: telemóvel vs tlm), hifenização, etc, leva a uma grande probabilidade de ocorrência de erros. Para lidar com tais situações, foram consideradas várias soluções para a correção dos erros ortográficos das palavras, nomeadamente a:

1. Utilização de um corretor ortográfico já existente.

2. Implementação de um corretor ortográfico (automático).

3. Correção manual de palavras relevantes.

A primeira solução seria a ideal em termos do tempo necessário para o desenvolvimento, dado que seria utilizado um corretor ortográfico já existente. A segunda solução seria a mais interessante em termos práticos. No entanto, estas duas soluções apresentam cada uma um problema fundamental. Devido ao idioma da linguagem utilizado no conjunto de dados (Português) não foi encontrado qualquer corretor ortográfico de fácil integração no processo de pré-processamento e cuja qualidade dos resultados fosse aceitável (nenhum corretor ortográfico existente é completamente fiável e preciso, podendo conter erros nas correções como a não deteção de um erro ortográfico ou a incorreta correção de um erro existente). A construção de um corretor ortográfico cuja qualidade dos resultados seja aceitável é uma tarefa cujos recursos necessários (dados de treino e teste para modelos de aprendizagem automática, regras gramaticais, etc) e tempo de implementação vão para além do tema da presente dissertação, podendo por si só ser um possível projeto de dissertação. Dadas estas dificuldades foi utilizada a terceira solução para correção dos erros ortográficos. Para esta solução baseada na correção manual foi implementado um dicionário com a correspondência entre palavras típicas com erros ortográficos e a palavra natural. O conjunto de palavras com erros ortográficos utilizadas foi sustentado pela análise das palavras mais e menos frequentes aqui apresentada e ainda pela análise natural dos comentários ao longo do desenvolvimento das soluções propostas. O dicionário desenvolvido possui, então, informação sobre os erros mais comuns e as palavras mais suscetíveis à ocorrência de erros – a tabela 3.9 mostra algumas das entradas do dicionário.

[4]Entenda-se palavra natural por a palavra sobre a qual ocorreram erros ortográficos, ou seja, a palavra correta

Erro	Palavra Natural
agradavel	agradável
camara	câmera
cm	com
confortavel	confortável
cumple	cumpre
chelente	excelente
fragil	frágil
k	que
util	útil

Tabela 3.9: Dicionário de erros comuns (extrato).

Remoção de Sinais de Pontuação, Caracteres Especiais, Números *Stop Words*

Como foi visto na figura 3.9 do ponto de análise anterior, algumas das palavras mais frequentes são sinais de pontuação. Para além das palavras em si e dos sinais de pontuação estão também presentes alguns números nos comentários. Os sentimentos são expressos pela utilização de palavras de sentimentos que são os maiores indicadores da polaridade de um documento, estas palavras que podem ser classificadas como positivas (ex.: perfeito, bom) ou negativas (ex.: mau, imperfeito) são, portanto, o ponto de análise essencial levando a que qualquer outra informação presente no texto de análise seja frequentemente considerada como ruído, diminuindo a qualidade dos dados e, por conseguinte, a capacidade de utilização por parte de modelos de modo a que a qualidade dos resultados seja aceitável. O conjunto de transformações a aplicar sobre um comentário deve então aproximar o mais possível esse comentário do conjunto de palavras de sentimento nele contidas, excluindo qualquer outra informação que não seja útil para o reconhecimento da polaridade. Neste sentido foram separados de cada comentário os sinais de pontuação, os números e as palavras em si para análise.

As figuras 3.10 e 3.11 mostram a distribuição dos 50 sinais de pontuação e números mais frequentes, respetivamente. É possível observar que nem todos são sinais de pontuação, a presença de caracteres especiais e conjuntos de caracteres especiais como sinais matemáticos e emojis estão também incluídos nesta separação. Isto acontece também para os números, onde a junção de números e letras é também considerado. De uma forma geral, a separação dos sinais de pontuação e dos números das palavras é efetuada de modo a que qualquer caracter especial ou conjunto de caracteres especiais e número ou conjunção de números com caracteres especiais ou letras seja excluído, obtendo assim apenas o conjunto de todas as palavras individuais. Esta exclusão é possível dado que estes componentes não possuem qualquer sentido ou influência

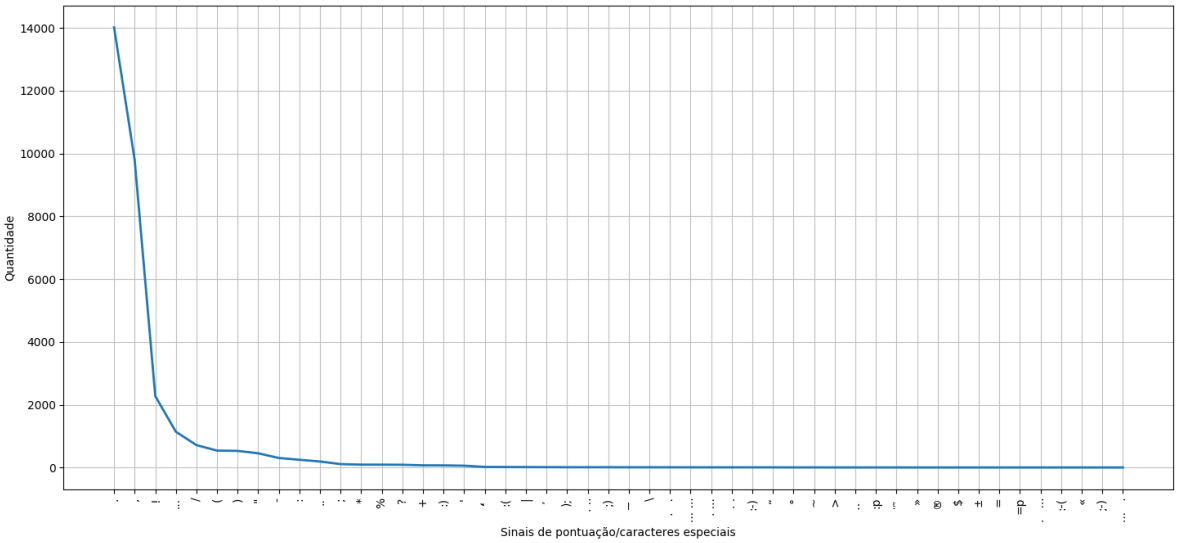


Figura 3.10: Distribuição dos 50 sinais de pontuação/caracteres especiais mais frequentes dos comentários.

.	,	!	...	/	(	)	"	-	:
..	;	*	%	?	+	:)	’	€	:(
|	\	);	. ..	;)	–	\\	. .	… …	. …
. .	:-)	’	º	~	>	. . .	:p	TM	»
®	$	±	=	=p	. …	:-(	«	;-)	… .

Tabela 3.10: 50 sinais de pontuação/caracteres especiais mais frequentes dos comentários.

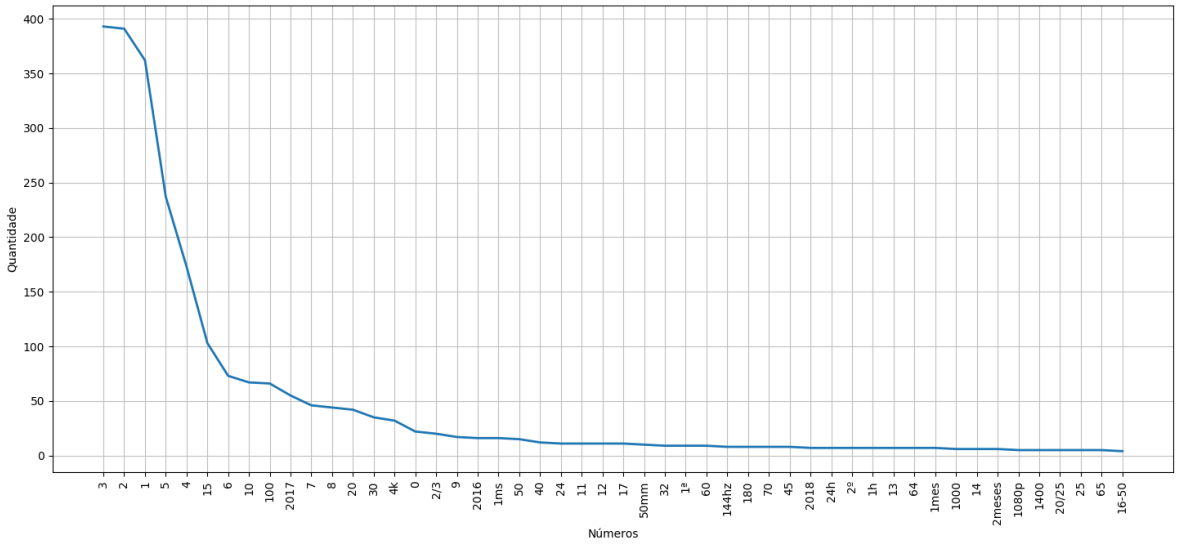


Figura 3.11: Distribuição dos 50 números mais frequentes dos comentários.

3	2	1	5	4	15	6	10	100	2017
7	8	20	30	4k	0	2/3	9	2016	1ms
50	40	24	11	12	17	50mm	32	1ª	60
144hz	180	70	45	2018	24h	2º	1h	13	64
1mes	1000	14	2meses	1080p	1400	20/25	25	65	16-50

Tabela 3.11: 50 números mais frequentes dos comentários.

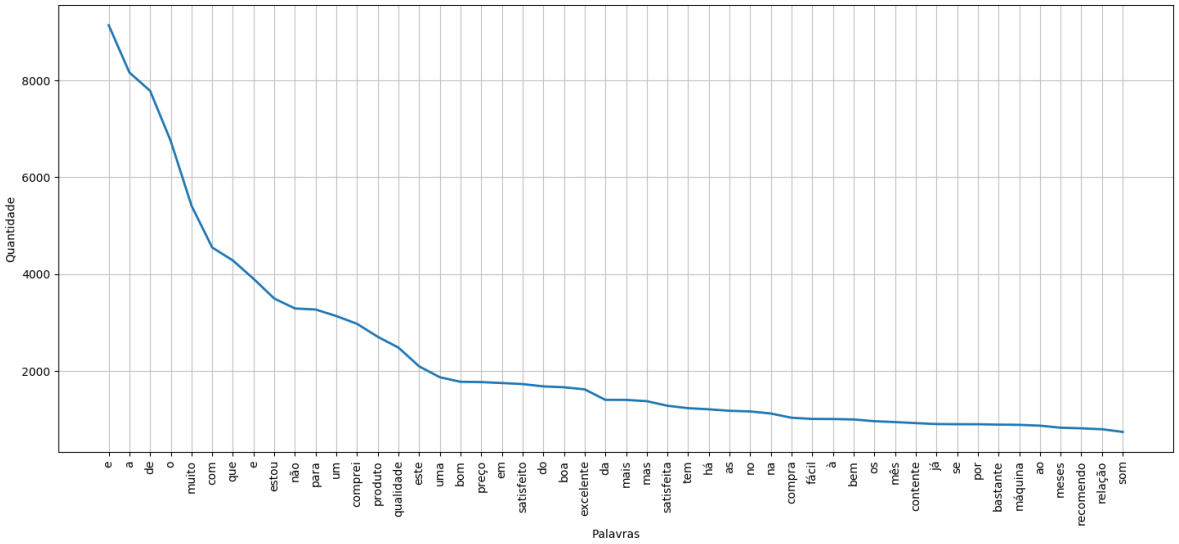


Figura 3.12: Distribuição das 50 palavras mais frequentes dos comentários.

e	a	de	o	muito	com	que	é	estou	não
para	um	comprei	produto	qualidade	este	uma	bom	preço	em
satisfeito	do	boa	excelente	da	mais	mas	satisfeita	tem	há
as	no	na	compra	fácil	à	bem	os	mês	contente
já	se	por	bastante	máquina	ao	meses	recomendo	relação	som

Tabela 3.12: 50 palavras mais frequentes dos comentários.

sobre a polaridade dos comentários. Note-se que é efetuada uma exceção na exclusão de emojis que ao contrário dos restantes componentes expressam um determinado sentimento. A figura 3.12 mostra o resultado desta exclusão, que permitiu obter, assim, apenas as palavras.

A distribuição da figura 3.12 mostra que muitas das palavras mais frequentes são *stop words*.

- ***Stop Words*** – conjunto de palavras que são excluídas durante o processamento de linguagem natural. Apesar de não existir uma definição que indique exatamente quais as palavras que devem pertencer a este conjunto, usualmente são consideradas as palavras mais frequentes do idioma da linguagem dos dados de análise.

Dependendo do problema de análise, a remoção de *stop words* pode ou não ser efetuada. A decisão de remover qualquer palavra do conjunto de dados de análise deve ser efetuada tendo em conta o impacto dessa palavra sobre os dados, se é importante para o processamento e correto resultado final. A tabela 3.13 apresenta algumas (50) das *stop words* mais frequentes da língua portuguesa. Como seria de esperar, palavras como *"de"*, *"a"*, *"o"*, *"e"*, *"não"*, *"mas"* fazem parte do conjunto de *stop words* dada as suas elevadas frequências em qualquer documento. No sentido da análise de sentimentos, estas palavras não expressam qualquer sentimento ou sentido de opinião logo podem ser removidas das restantes palavras, diminuindo a quantidade de ruído na utilização pelos modelos, especialmente de aprendizagem automática. Neste sentido, foi removido o conjunto de *stop words* portuguesas, consideradas na biblioteca NLTK [7], constituído por 203 palavras distintas (algumas já apresentadas na tabela abaixo).

de	a	o	que	e	do	da	em	um	para
com	não	uma	os	no	se	na	por	mais	as
dos	como	mas	ao	ele	das	à	seu	sua	ou
quando	muito	nos	já	eu	também	são	pelo	pela	até
isso	ela	entre	depois	sem	mesmo	aos	seus	quem	nas

Tabela 3.13: Extrato do conjunto de *stop words* utilizadas.

Apesar de não ser expresso qualquer sentimento pelas *stop words*, na verdade nem todas foram removidas dado que algumas são de extrema importância na forma como modificam ou suportam uma palavra ou frase que exprime um sentimento. A melhor palavra para exemplificar esta influência que certas *stop words* podem exercer sobre os sentimentos é a palavra "*não*". Considere-se o exemplo abaixo:

Exemplo 3: Considere-se os possíveis comentários:

C1: *Gostei do produto*

C2: *Não gostei do produto*

No exemplo acima, a palavra "*não*" é uma *stop word* e tem um dos maiores papeis em análise de sentimentos – capacidade de modificar um sentimento. A remoção da palavra "*não*" do comentário 2 faz com que o comentário perca o sentido, o papel desta palavra é modificar a polaridade da palavra de sentimento "*gostei*", normalmente positiva, tornando todo o comentário negativo (não existem mais palavras de sentimento). O conjunto de *stop words* não removidas segue este princípio de influência sobre uma dada palavra de sentimento ou frase e até de conector entre palavras e frases que sem ele o sentido dos sentimentos expressos pode ser perdido. A *stop word* "*mas*" representa este efeito onde, no exemplo 4, a presença desta palavra tem um efeito conector entre as duas partes principais do comentário e indica ainda a polaridade da segunda parte, que sem ela poderia ser facilmente perdida. O efeito da

palavra "*mas*" é de contraste, inserida no conjunto de palavras de contraste (*but-clauses*), já apresentadas no capítulo 2 no nível de classificação entidade/aspeto. O efeito destas palavras de contraste vai ser ainda aplicado com mais detalhe posteriormente na implementação de modelos baseados em dicionários.

Exemplo 4: Considere-se o possível comentário:

C1: *Não gostei do telemóvel mas a bateria é longa*

O conjunto de *stop words* não removidas é ainda diferente consoante a solução utilizada, soluções baseadas em dicionários necessitam em certos casos de uma maior estrutura dos comentários onde a presença de certas formas verbais, maioritariamente expressas pelos sufixos das palavras, pode auxiliar na forma de tratamento dos comentários e na identificação de sentimentos – posteriormente, apresentaremos alguma informação mais detalhada sobre a importância da não remoção de certas *stop words* nos modelos baseados em dicionários é apresentada no respetivo capítulo. A tabela 3.14 mostra o conjunto total de *stop words* que não foram removidas para cada tipo de solução.

	Palavras	Palavras extra
Aprendizagem Automática	não / mas / muito / são / sem / mesmo / nem / mais / seria / fosse / só / um	–
Dicionários		isso / disto / aquilo / este / estes / esta / estas / esse / esses / essa / essas

Tabela 3.14: *Stop words* não removidas em cada solução (aprendizagem automática e dicionários).

A análise e tratamento dos dados efetuado neste ponto, maioritariamente a remoção de sinais de pontuação ou conjuntos de caracteres especiais, números e palavras que contenham números e *stop words*, permitiu reduzir o tamanho do vocabulário diminuindo a quantidade de ruído nos dados. A figura 3.13 e a tabela 3.15 mostram a distribuição das 50 palavras mais frequentes após os tratamentos aqui apresentados. A análise da distribuição mostra que, ao contrário das distribuições anteriormente apresentadas, as palavras mais frequentes começam a ser caracteristicamente palavras de sentimento como é o caso das palavras *"bom"*, *"satisfeito"*, *"boa"*, *"excelente"*, etc. As três palavras mais frequentes são agora *"muito"*, *"é"* e *"não"*. A grande frequência destas palavras é sustentadamente justificável pelo facto de serem palavras muito comuns em qualquer documento e mais especificamente no contexto de análise de sentimentos. Por exemplo, *"muito"* é comummente utilizado em comentários de qualquer polaridade, tal como se pode ver no exemplo 5.

Exemplo 5: Considere-se os possíveis comentários:

C1: *Gostei muito do produto*

C2: *Não gostei muito do produto*

O exemplo acima serve também para justificar a frequência da palavra *"não"* que para além de poder ser utilizada como negação de palavras de sentimento pode também ser utilizada livremente nos comentários. Por último, a palavra *"é"* (forma verbal do verbo ser/estar) é frequentemente utilizada como forma de referenciar um produto ou aspeto de um produto.

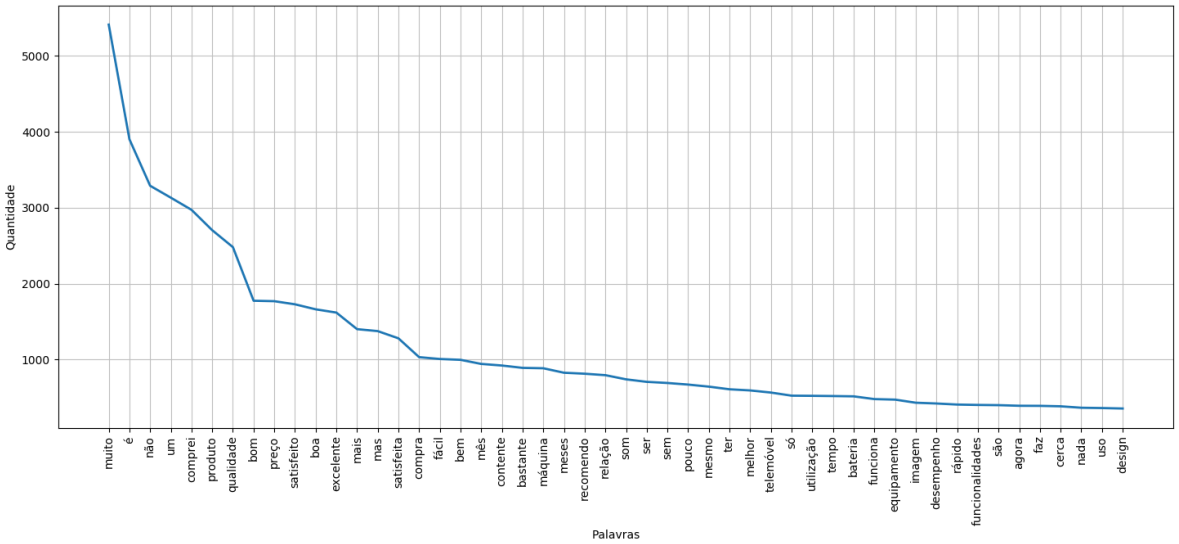


Figura 3.13: Distribuição das 50 palavras mais frequentes dos comentários.

muito	é	não	um	comprei	produto	qualidade	bom	preço
satisfeito	boa	excelente	mais	mas	satisfeita	compra	fácil	bem
mês	contente	bastante	máquina	meses	recomendo	relação	som	ser
sem	pouco	mesmo	ter	melhor	telemóvel	só	utilização	tempo
bateria	funciona	equipamento	imagem	desempenho	rápido	funcionalidades	são	agora
faz	cerca	nada	uso	design				

Tabela 3.15: 50 palavras mais frequentes dos comentários.

Análise das Palavras por Comprimento

A análise e transformações até agora efetuadas foram aplicadas globalmente aos comentários e palavras e permitiram uniformizar a qualidade e formato dos dados. Com a remoção de *stop words* conseguiu-se aproximar o vocabulário do conjunto de palavras de sentimento. Con-

tudo, existem ainda palavras cuja relevância para análise de sentimentos não é significativa, acrescentando ruído aos dados. Dada a impossibilidade de análise de todas as palavras de forma individual, foi necessário analisar as palavras segundo uma propriedade comum que as relacione e de maior detalhe, neste caso, o comprimento de cada uma. A figura 3.14 mostra a distribuição do comprimento das palavras dos comentários.

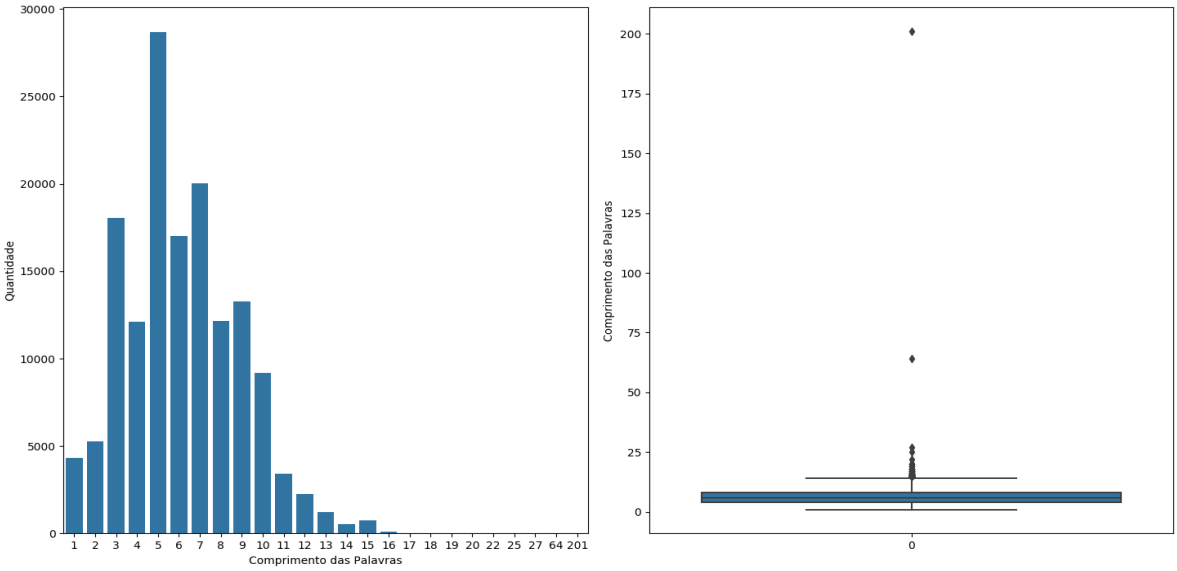


Figura 3.14: Distribuição do comprimento das palavras.

As palavras mais pequenas têm um comprimento igual a 1 e as maiores têm um comprimento igual a 201. O comprimento mais frequente é igual a 5. Existem ainda palavras cujo comprimento é considerado um *outlier*. Neste sentido, foram considerados *outliers* todos os comprimentos iguais ou superiores a 16. Estes valores definem um intervalo de comprimentos das palavras sobre o qual vai ser efetuada a análise.

A análise das palavras mais pequenas foi efetuada entre os comprimentos 1 e 6 dado que o comprimento mais frequente é 5.

- **Comprimento igual a 1:**

é	i	s	q	x	d	g	è	c	p	b	_	v	m
l	f	ó	h	ê	u	w	y	j	ú	ª	å	t	z

Tabela 3.16: Palavras com comprimento igual a 1.

Existem 28 palavras distintas com comprimento igual a 1, apresentadas na tabela 3.16. Em termos de análise de sentimentos estas palavras não são relevantes e foram,

portanto, removidas.

- **Comprimento igual a 2:**
  A análise das palavras mais frequentes com este comprimento mostrou que existem algumas delas que são relevantes para o processo de análise de sentimentos, como as palavras *"má"* e *"ok"*, impossibilitando a remoção total de todas as palavras com comprimento igual a dois. No sentido de conseguir remover algumas das palavras com este comprimento foi efetuada uma análise àquelas que ocorrem apenas uma vez no vocabulário. Estas palavras são usualmente denominadas por *hápaxes*.

éo	hc	éa	gm	gh	le	lb	yw	lp	ti
dn	dx	i7	dr	ds	dp	qq	qr	qu	qt
t1	qe	kl	el	ei	ee	j5	j7	w7	rt
rf	rb	ju	bn	bt	z8	js	om	zm	tc
câ	xl	f1	xc	ví	ce	xx	e4	xp	pr

Tabela 3.17: 50 hápaxes de comprimento igual a 2.

  A tabela 3.17 mostra 50 dos 80 hápaxes de comprimento igual a dois. Nenhuma destas palavras é relevante para análise de sentimentos, logo, foram removidas do vocabulário.

- **Comprimento igual a 3, 4 e 5:**
  A análise das palavras com estes comprimentos segue um padrão semelhante à análise de comprimento igual a dois. Foram removidos os hápaxes destes comprimentos de acordo com a tabela 3.18.

Comprimento	Quantidade Removida (nº hápaxes)
3	148
4	293
5	462

Tabela 3.18: Quantidade de hápaxes removidos para os comprimentos 3, 4 e 5.

- **Comprimento igual a 6:**
  Tal como nos comprimentos anteriores, as palavras mais frequentes expressam sentimentos e, como tal, impossibilitam a sua remoção do vocabulário. A análise dos hápaxes, ao contrário dos comprimentos anteriores, contém, também, palavras relevantes para análise de sentimentos como *"amámos"*, *"piorou"*, o que impossibilita a sua remoção.

A não remoção de qualquer palavra com comprimento igual a seis, e dado que este comprimento

é imediatamente a seguir ao comprimento de maior frequência, indica que se atingiu um bom limite de análise para as palavras mais pequenas. A análise das palavras de comprimento maior, igual ou superior a 16, mostrou que apenas as palavras de comprimento igual ou superior a 20 podem ser removidas dado que algumas das palavras de comprimentos inferiores expressam sentimentos. A tabela 3.19 mostra um extrato do conjunto das 78 palavras de comprimento superior.

Comprimento	Palavras
16	recomendadíssimo / profissionalismo / descontentamento / convenientemente
17	satisfatoriamente / espectacularmente / independentemente / incomparavelmente
18	surpreendentemente / significativamente / excelenteexcelente / satisfeitoobrigado
19	impressionantemente / transportardobra-se / extraordináriamente
20	kkkkkkkkkkkkkkkkkkk / recomendadrecomendad / profissionaisdestaco
22	playmemoriescameraapps
25	aaaaaaaaaaaaaaaaaaaaaaaaa
27	xxxxxxxxxxxxxxxxxxxxxxxxxxx
64	peperfeitorfeitoperfeitoperfeitoperfeitoperfeitoperfeitoperfeito
201	booooooooooooooooooooiioiioioooooioooooooooooooooommmmmmmmmmm...

Tabela 3.19: Palavras de comprimento superior (extrato).

Análise das Palavras por Polaridade e Correlação das Palavras por Polaridade

O ponto de análise anterior representou as últimas transformações aplicadas sobre o conjunto de dados durante o pré-processamento. Todas as transformações aplicadas têm o objetivo de uniformizar os dados pela remoção de informação irrelevante para análise de sentimentos. Agora que o conjunto de dados sofreu todas as transformações, está preparado para ser utilizado por modelos e soluções de análise de sentimentos. Porém, antes disso, é importante analisar o resultado da aplicação das transformações. A análise das palavras por polaridade para além de permitir validar as transformações, permite também perceber a forma como as palavras são utilizadas em comentários positivos e negativos. As figuras 3.15 e 3.16 e as tabelas 3.20 e 3.21 mostram, respetivamente, a distribuição das 50 palavras mais frequentes dos comentários positivos e negativos. É possível observar que os comentários positivos utilizam frequentemente palavras de sentimento positivas, tal como seria de esperar. A palavra de sentimento positiva mais frequente nos comentários é *"bom"* seguida de *"satisfeito"*, *"excelente"*, *"boa"*, etc. Qualquer outra palavra que não expresse diretamente um sentimento, tanto pode ser uma palavra de sentimento dependente do contexto como uma palavra comum que serve de conector entre as várias palavras, de modo a manter uma estrutura do comentário

potencialmente relevante para um processo de análise de sentimentos (ex.: gostei vs gostei muito).

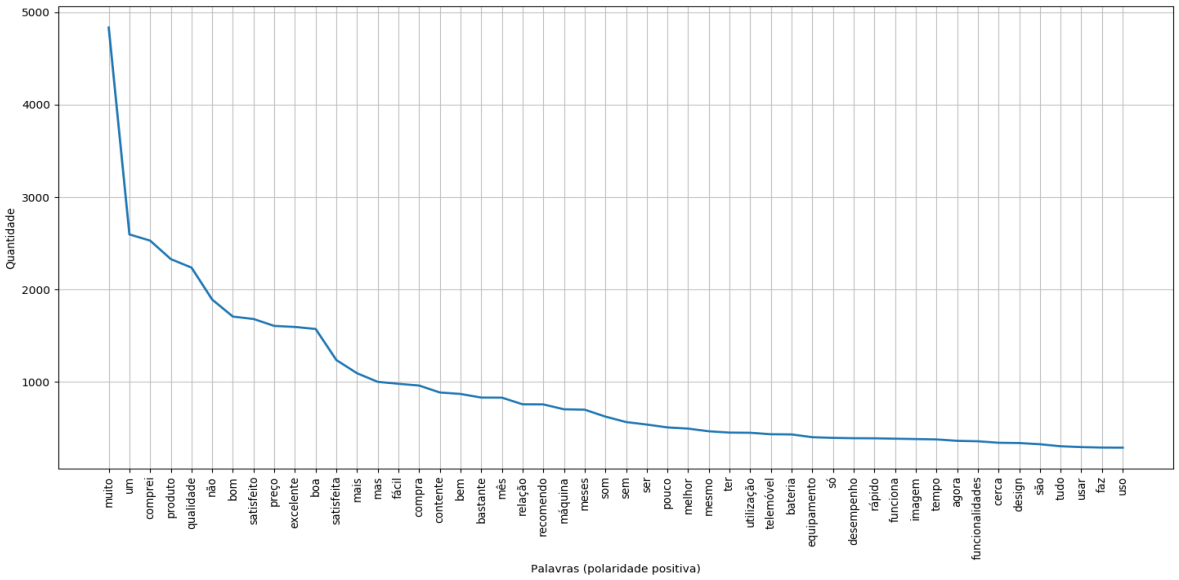


Figura 3.15: Distribuição das 50 palavras mais frequentes dos comentários positivos.

muito	um	comprei	produto	qualidade	não	bom	satisfeito	preço	excelente
boa	satisfeita	mais	mas	fácil	compra	contente	bem	bastante	mês
relação	recomendo	máquina	meses	som	sem	ser	pouco	melhor	mesmo
ter	utilização	telemóvel	bateria	equipamento	só	desempenho	rápido	funciona	imagem
tempo	agora	funcionalidades	cerca	design	são	tudo	usar	faz	uso

Tabela 3.20: 50 palavras mais frequentes dos comentários positivos.

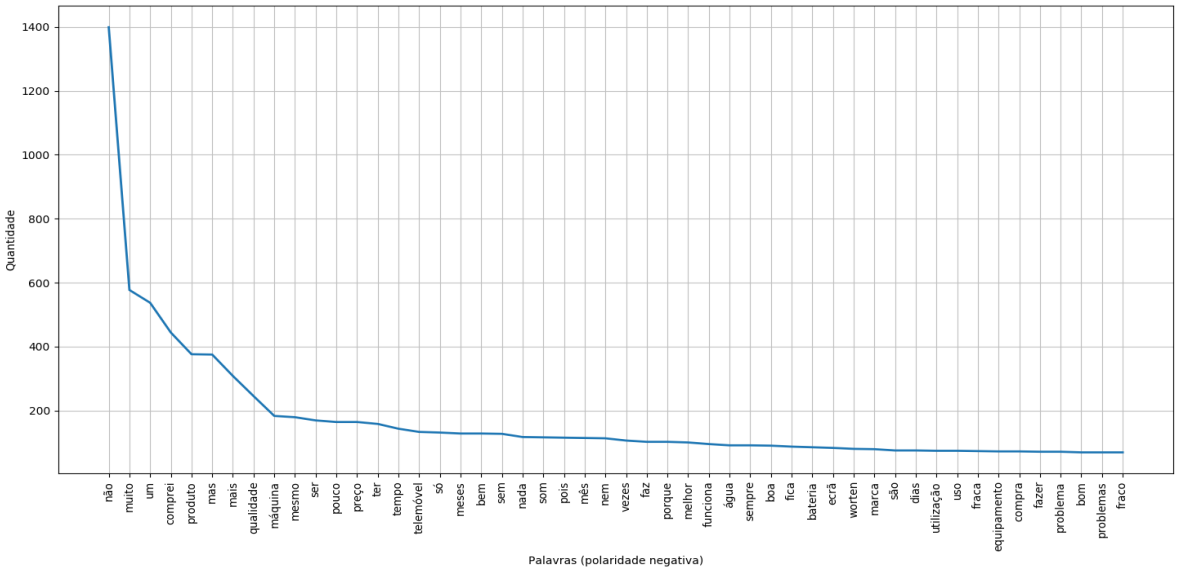


Figura 3.16: Distribuição das 50 palavras mais frequentes dos comentários negativos.

não	muito	um	comprei	produto	mas	mais	qualidade	máquina	mesmo
ser	pouco	preço	ter	tempo	telemóvel	só	meses	bem	sem
nada	som	pois	mês	nem	vezes	faz	porque	melhor	funciona
água	sempre	boa	fica	bateria	ecrã	worten	marca	são	dias
utilização	uso	fraca	equipamento	compra	fazer	problema	bom	problemas	fraco

Tabela 3.21: 50 palavras mais frequentes dos comentários negativos.

Ao contrário do que acontece com os comentários positivos, os comentários negativos não utilizam frequentemente palavras de sentimento negativas, tal como é possível ver pela figura 3.16 e tabela 3.21. Ocorrem também até palavras de sentimento positivas como *"boa"* e *"bom"*. Uma observação mais cuidada sobre as palavras mais frequentes desta polaridade permite justificar e compreender esta falta de palavras de sentimento negativas mais frequentes – a palavra mais frequente é *"não"* e considere-se ainda as palavras *"pouco"*, *"nada"*, *"melhor"*, cuja frequência é relativamente elevada, estas palavras são modificadores de sentimentos (*sentiment shifters*) que alteram a polaridade das palavras de sentimento a elas associadas. Atente-se ao exemplo 6.

Exemplo 6: Considere-se os possíveis comentários:

C1: *Não gostei*

C2: *Gostei pouco*

C3: *Nada satisfeito*

C4: *Podia ser melhor*

Os comentários do exemplo acima são todos negativos, mas nenhum deles utiliza palavras de sentimento negativas. Pelo contrário, os três primeiros comentários possuem palavras de sentimento positivas. Os comentários tornam-se, então, negativos pela utilização das palavras modificadoras de sentimentos. Estas observações sugerem, portanto, que uma grande parte da forma como os comentários negativos são construídos é pela negação de palavras de sentimento positivas.

Uma outra análise relevante para perceber como os comentários são escritos no sentido da forma como as palavras se relacionam entre si e que pode ser utilizada para validar a análise acima efetuada é a correlação das palavras. Considera-se aqui a correlação de uma palavra com um conjunto de palavras como a quantidade de vezes que essa palavra ocorre junto das restantes. Estas quantidades são representadas numa matriz e podem ser facilmente desenhadas num grafo. As figuras 3.17 e 3.18 mostram os grafos da correlação entre as 100 palavras mais frequentes em cada polaridade.

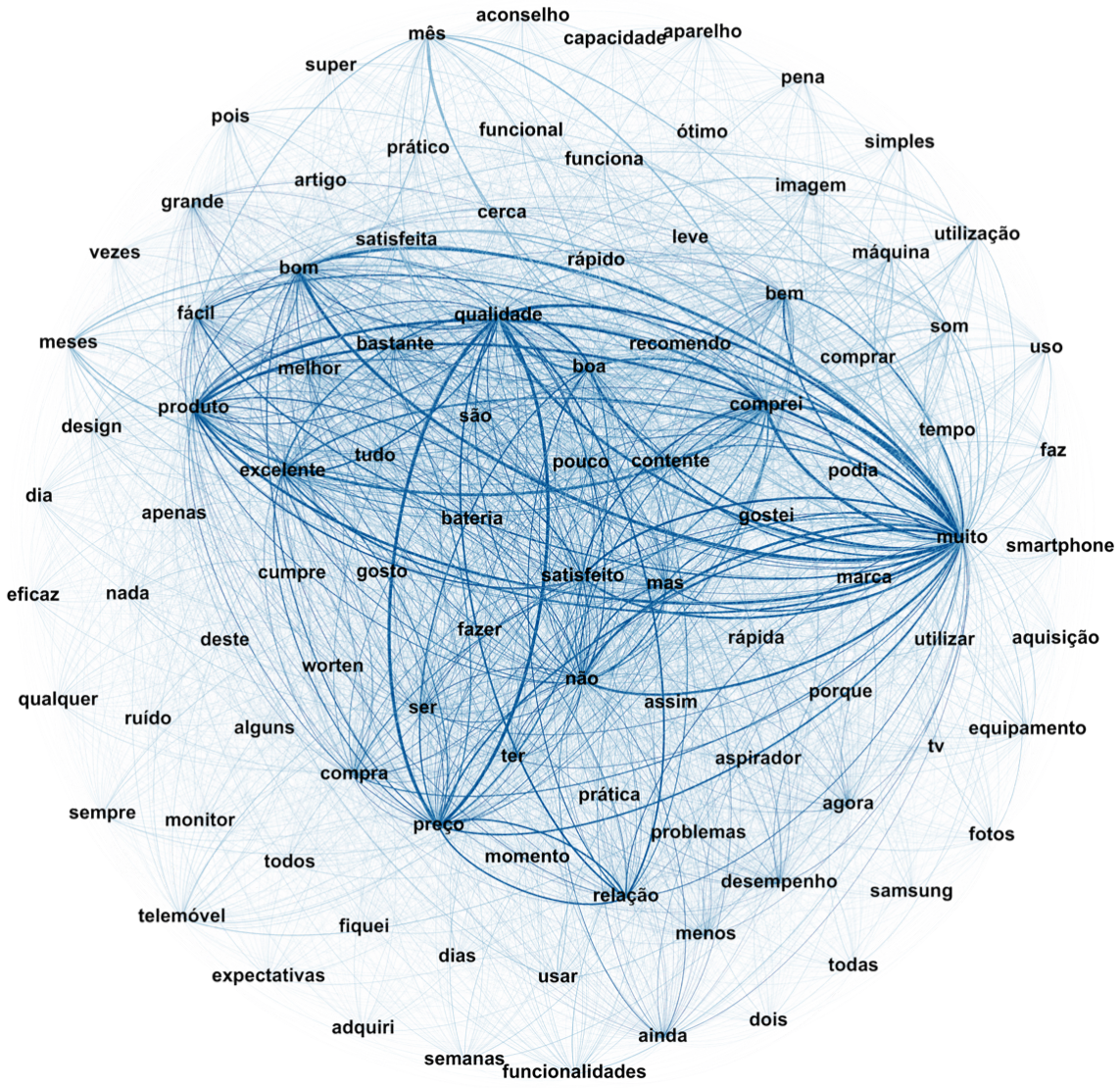


Figura 3.17: Correlação entre as 100 palavras mais frequentes dos comentários positivos.

A representação em grafo permite observar facilmente os pequenos clusters que foram estabelecidos entre as maiores correlações das palavras. O grafo da correlação entre as palavras de polaridade positiva (figura 3.17) mostra a correlação entre várias palavras como:

- *boa relação preço qualidade*
- *produto muito bom*
- *qualidade excelente*
- *muito satisfeito produto*

Todas estas correlações entre as palavras formam comentários positivos, como seria de esperar, e permitem perceber como estes comentários são frequentemente escritos.

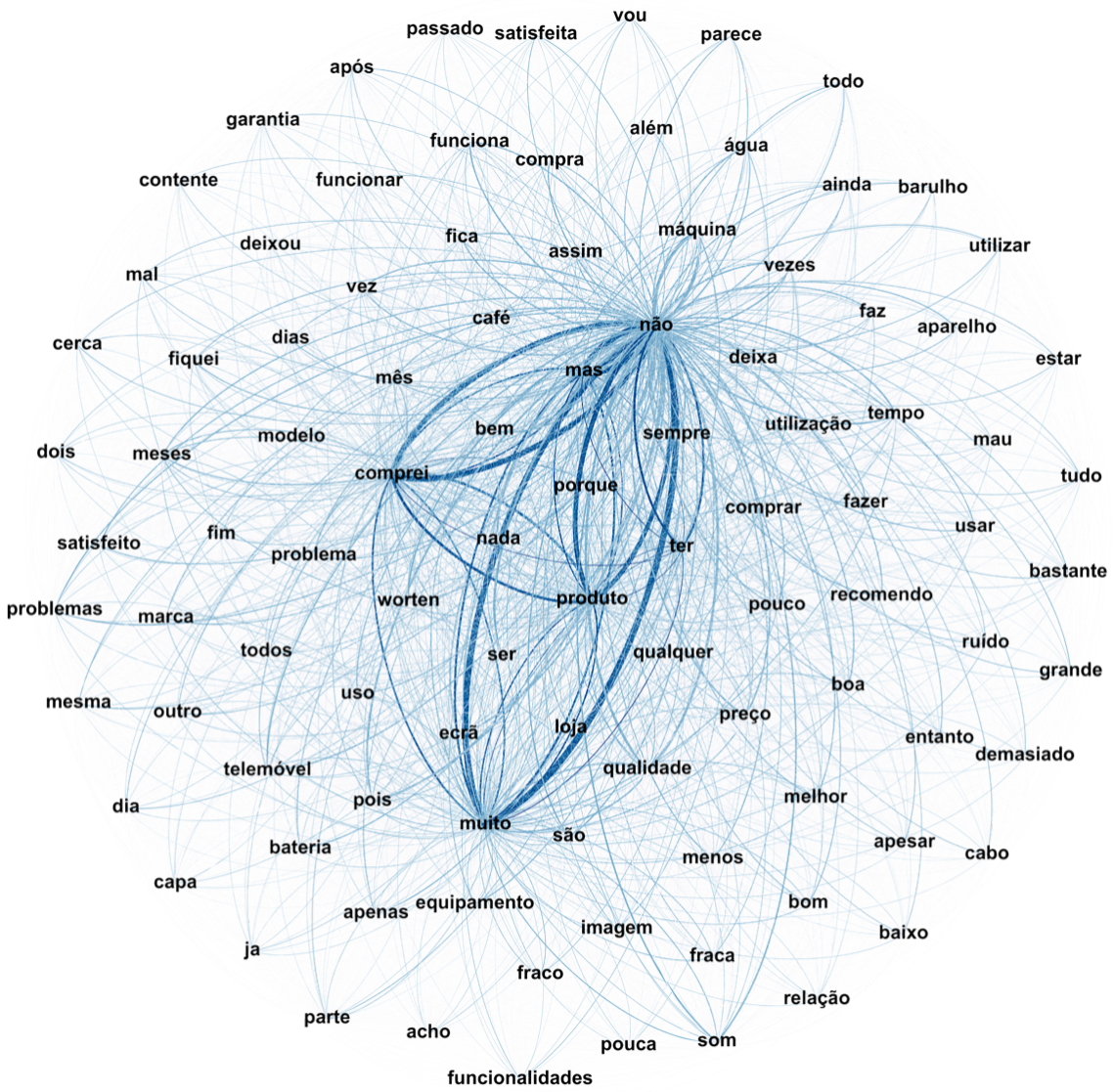


Figura 3.18: Correlação entre as 100 palavras mais frequentes dos comentários negativos.

A correlação entre as palavras de polaridade negativa (figura 3.18) permite confirmar a análise realizada anteriormente sobre a forma como os comentários negativos são frequentemente construídos – pela negação de palavras de sentimento positivas – tal como é possível observar no grafo através da frequência elevada que a palavra *"não"* possui com as restantes palavras.

3.3.3 Pré-Processamento – Resumo e Conclusões

A fase de pré-processamento permite analisar e aplicar uma série de transformações sobre um dado conjunto de dados. A escolha de cada transformação é usualmente determinada pelos dados e pelo resultado de cada uma das transformações antecedentes, com o objetivo de uniformizar o conjunto de dados, pela possibilidade de acréscimo de novos atributos (para a análise efetuada e para o conjunto de dados utilizados não foi relevante) e pela remoção de informação não relevante, de modo a reduzir o tamanho do vocabulário, o tanto quanto possível, para que o conjunto de dados final seja constituído maioritariamente por palavras de sentimento e outras palavras necessárias, de forma a garantir uma estrutura mínima para cada comentário reduzindo assim a quantidade de ruído nos dados.

A análise efetuada depois da aplicação das transformações permitiu avaliar o efeito da sua aplicação, pela forma como os comentários são construídos ou escritos em cada polaridade. Resta, portanto, verificar a redução sobre o vocabulário provocada pelas transformações aplicadas.

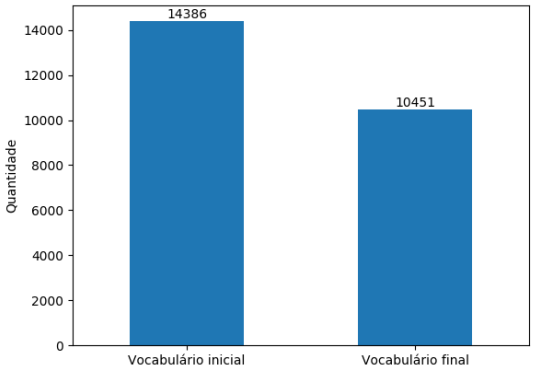


Figura 3.19: Tamanho do vocabulário inicial e final após aplicação das transformações.

A figura 3.19 mostra a redução do vocabulário conseguida depois da aplicação das transformações. Conseguiu-se uma redução do vocabulário igual a 3935 palavras (potencialmente), não relevantes, cuja presença origina apenas ruído nos dados, o que dificulta o seu processamento e utilização por modelos de aprendizagem automática. Note-se que o conjunto de transformações aplicado não é de todo um referência exata para qualquer pré-processamento a ser aplicado numa outra solução de análise de sentimentos ou de processamento de linguagem natural. As transformações aqui aplicadas foram de acordo com o tipo e qualidade dos dados. Outras transformações podiam ser também aplicadas de modo a uniformizar ainda mais o conjunto de dados. A escolha de um bom ponto de paragem no pré-processamento dos dados é, portanto, importante, devendo o seu resultado ser validado pelos modelos que utilizam os dados, para se perceber se é necessário um pré-processamento mais detalhado sobre os da-

dos, de forma a que seja possível atingir o nível de precisão desejado com os modelos. Como forma de conclusão, considere-se o exemplo abaixo representativo da aplicação de todas as transformações do pré-processamento sobre um comentário do conjunto de dados.

Exemplo 7: Considere-se o comentário do conjunto de dados:

C1: *Comprem,comprei e já é o melhor telemovel que tive.grande máquina com muita qualidade só aquece um bocado.*

Transformação	Comentário
Separação das Palavras	*Comprem, comprei e já é o melhor telemovel que tive. grande máquina com muita qualidade só aquece um bocado.*
Minúsculas	*comprem, comprei e já é o melhor telemovel que tive. grande máquina com muita qualidade só aquece um bocado.*
Tokens	*['comprem', ',', 'comprei', 'e', 'já', 'é', 'o', 'melhor', 'telemovel', 'que', 'tive', '.', 'grande', 'máquina', 'com', 'muita', 'qualidade', 'só', 'aquece', 'um', 'bocado', '.']*
Corretor Ortográfico	*['comprem', ',', 'comprei', 'e', 'já', 'é', 'o', 'melhor', 'telemóvel', 'que', 'tive', '.', 'grande', 'máquina', 'com', 'muita', 'qualidade', 'só', 'aquece', 'um', 'bocado', '.']*
Remoção Pontuação & Números	*['comprem', 'comprei', 'e', 'já', 'é', 'o', 'melhor', 'telemóvel', 'que', 'tive', 'grande', 'máquina', 'com', 'muita', 'qualidade', 'só', 'aquece', 'um', 'bocado']*
Remoção Stop Words	*['comprem', 'comprei', 'é', 'melhor', 'telemóvel', 'grande', 'máquina', 'muita', 'qualidade', 'só', 'aquece', 'um', 'bocado']*
Remoção Palavras Pequenas e Grandes	*['comprem', 'comprei', 'melhor', 'telemóvel', 'grande', 'máquina', 'muita', 'qualidade', 'só', 'aquece', 'um', 'bocado']*

Tabela 3.22: Aplicação das transformações do pré-processamento sobre um comentário do conjunto de dados.

4 A Construção de Recursos

Os recursos que são utilizados nas soluções de análise de sentimentos podem ser vários. A qualidade dos resultados de qualquer solução é dependente da qualidade dos constituintes dessa solução. Um qualquer recurso utilizado, que produza resultados que sirvam de input para um outro recurso e cuja precisão e qualidade dos mesmos não seja suficientemente boa, pode comprometer todo o sistema, este problema pode ser visto como um efeito em cascata, em que a qualidade dos resultados de cada recurso utilizado vai diminuindo de recurso para recurso. Assim, considere-se que um recurso é uma qualquer solução que recebe um conjunto de dados como input, processa os mesmos e devolve um output que pode ser o resultado final de todo o sistema ou pode ser utilizado como input por um outro recurso. De entre os recursos utilizados é importante salientar e apresentar detalhadamente a implementação e operacionalidade de dois deles:

- Léxico/Dicionário de Sentimentos
- POS Tagger

4.1 Léxico/Dicionário de Sentimentos

Este é o recurso desenvolvido com maior importância, uma vez que é o componente fundamental do modelo baseado em dicionários (capítulo 6). Existem várias formas de construção de um dicionário de sentimentos, algumas já apresentadas no capítulo 2. O léxico deste trabalho foi desenvolvido fazendo uso de recursos já existentes, nomeadamente, dois léxicos da língua portuguesa, nomeadamente:

- *SentiLex-PT01* (versão *flex*) [32]
- *Priberam Subjectivity Lexicon for Portuguese* [3]

De uma forma resumida, estes dois léxicos têm as seguintes características e são constituídos por:

- ***SentiLex-PT01***

  Este léxico é constituído por 25406 entradas. As palavras são apresentadas em género e número (esta foi uma característica essencial na utilização deste recurso dado que para cada palavra, a sua representação na forma singular, plural, masculino e feminino é também conhecida). Abaixo apresenta-se algumas entradas deste léxico:

  bonita,bonito.PoS=Adj;GN=fs;TG=HUM;POL=1;ANOT=MAN
  bonitas,bonito.PoS=Adj;GN=fp;TG=HUM;POL=1;ANOT=MAN
  bonito,bonito.PoS=Adj;GN=ms;TG=HUM;POL=1;ANOT=MAN
  bonitos,bonito.PoS=Adj;GN=mp;TG=HUM;POL=1;ANOT=MAN

  Para cada palavra do léxico está associado um conjunto de informação – *POS*, *GN*, *POL*, *TG* e *ANOT*. No sentido de desenvolver um novo léxico para análise de sentimentos, apenas é importante a informação: *POS* – Partes do Discurso e *POL* – Polaridade. Todas as palavras deste léxico são adjetivos e a polaridade das palavras pode ser positiva (1), negativa (-1) ou neutra (0).

- ***Priberam Subjectivity Lexicon***

  Este léxico é constituído por 9945 entradas. Abaixo apresenta-se algumas entradas deste léxico:

  gostar V strongsubj positive
  obscuro A weaksubj negative

  A informação associada a cada palavra é separada por tabulações e pela ordem: palavra, *POS*, intensidade e polaridade. Tal como no léxico anterior, apenas é relevante a informação de *POS* e polaridade (POL). Estas tags podem ter os valores de acordo com as tabelas abaixo.

POS Tag	Descrição
N	Nome/Substantivo
A	Adjetivo
D	Advérbio
V	Verbo
I	Interjeição
C	Conjunção
P	Preposição

Polaridade	Descrição
positive	positiva
negative	negativa
neutral	neutra
both	ambas (positiva e negativa)

O léxico de sentimentos foi construído com a junção destes dois recursos, sem a ocorrência

de repetições de palavras. Para tal foram efetuadas algumas transformações aos léxicos. De referir:

- A representação da polaridade do léxico *Priberam Subjectivity Lexicon* em -1 (negativo), 0 (neutro) e 1 (positivo) de modo a ter a mesma escala do outro léxico. Para a polaridade "ambas (both)" deste léxico foi efetuada uma revisão manual para definir a polaridade na nova escala.

- As tags das partes do discurso de ambos os léxicos foram modificadas de modo a terem a mesma representação das tags utilizadas no corpus *Projeto Floresta Sintáctica* [1], incluído na biblioteca NLTK. O motivo desta modificação para as *POS tags* deste corpus é por ser suportado pela biblioteca NLTK, utilizada em muito do processamento efetuado nos modelos desenvolvidos e para a implementação do *POS Tagger* (ver próxima secção).

O léxico construído é constituído por 28055 palavras distintas, anotadas com a respetiva parte do discurso e polaridade. Estas anotações permitem utilizar o léxico como um dicionário de sentimentos, dada a correspondência entre palavra – polaridade, e como forma de anotar palavras com as respetivas *POS tags* ou até ser utilizado como um recurso, de um modelo de aprendizagem automática para anotação de *POS tags*, dada a correspondência entre palavra – *POS tag*. A figura 4.1 apresenta informação relevante sobre a distribuição das anotações do léxico.

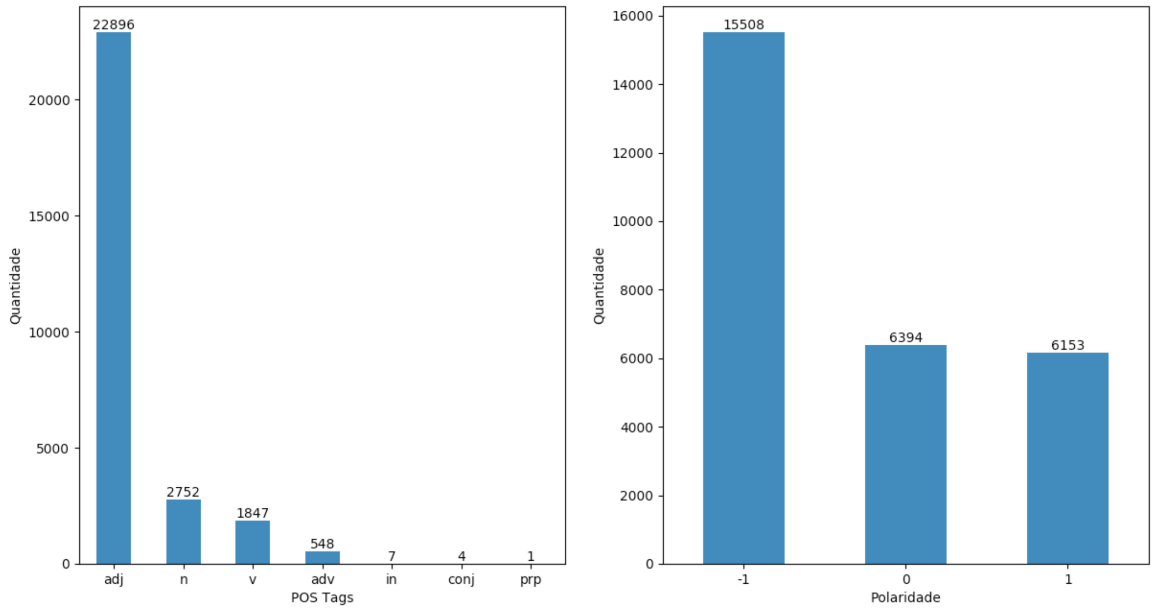


Figura 4.1: Distribuições das *POS tags* e da polaridade do léxico de sentimentos desenvolvido.

Apesar da utilização destes dois recursos para a construção do léxico, não é garantida a qualidade das anotações. Foi, portanto, efetuada uma revisão manual a uma grande parte do conjunto de palavras e respetivas anotações do léxico desenvolvido, nas quais foram corrigidas as polaridades a mais de 400 palavras. Para além disto, ao longo do desenvolvimento dos modelos, foi-se constatando a falta de palavras no léxico que são frequentemente utilizadas nos comentários e relevantes para determinar a polaridade dos mesmos. Neste sentido, foi acrescentado manualmente um conjunto de palavras e respetivas anotações, específicas para o contexto dos comentários do conjunto de dados, com mais de 350 palavras, frases ou expressões comuns (ex.: *"barato sai caro"* que é uma expressão caracteristicamente negativa mas se for tratada palavra a palavra ficaria neutra: barato (+1), caro (-1) dado que as polaridades individuais das palavras cancelam-se mutuamente).

4.2 *POS Tagger*

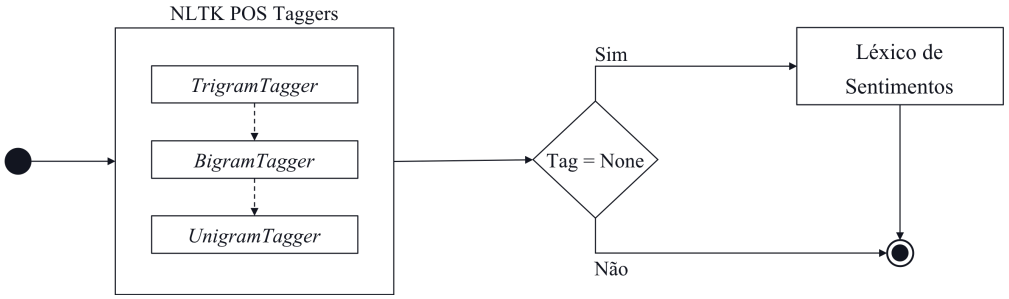


Figura 4.2: *POS Tagger* – Esquema de funcionamento.

O modelo de marcação de palavras de discurso desenvolvido utiliza três recursos principais:

- Biblioteca NLTK.
- Corpus *Floresta Sintática* (incluído na biblioteca NLTK).
- Léxico de sentimentos desenvolvido.

Um POS tagger processa um conjunto de palavras sequenciais e marca cada uma delas com a respetiva parte do discurso. No caso de uma palavra não ser reconhecida (não ser determinada a sua parte do discurso) ou é marcada como *None* ou com uma tag padrão.

A biblioteca NLTK possui várias soluções de marcação de partes do discurso, incluindo a capacidade de utilizar um corpus anotado para treino de um tagger. Podem ser treinados três tipos de taggers: *UnigramTagger*, *BigramTagger* e *TrigramTagger*. Como os próprios nomes

indicam, estes taggers utilizam conjuntos de n-gramas, de 1 a 3, respetivamente, para treino e marcação das tags. Para além disto, estes taggers podem ser treinados e utilizados em conjunto, especificando qual tagger utilizar como recuo de um outro tagger, caso o primeiro não consiga marcar uma palavra. Por exemplo, se a utilização de um *TrigramTagger* não conseguir marcar uma certa palavra, então o modelo treinado recua para um *BigramTagger* e para um *UnigramTagger* caso este também não consiga marcar a palavra.

Neste sentido, foi utilizado o corpus *Floresta Sintática* para treino de um modelo que faz uso de um *TrigramTagger*, recuando para um *BigramTagger* e deste para um *UnigramTagger*. No caso de o modelo não conseguir marcar uma palavra é-lhe atribuída a tag *None*. De modo a melhorar os resultados do modelo, pela marcação da maior quantidade possível de palavras, é utilizado o léxico de sentimentos desenvolvido para marcar cada palavra que tenha a tag *None*. No caso de a palavra não ser reconhecida pelo léxico então é aplicada a tag padrão "nome/substantivo", dado que é a parte do discurso mais frequente em qualquer documento da língua portuguesa, fazendo com que todas as palavras sejam marcadas. A figura 4.2 mostra este processo utilizado pelo *POS Tagger* desenvolvido.

Na secção anterior (desenvolvimento do léxico/dicionário de sentimentos) fez-se referência à transformação das tags das partes do discurso dos dois léxicos utilizados como recursos para as tags do corpus *Floresta Sintática* de modo a que fosse possível utilizar o léxico de sentimentos no *POS Tagger* dado que os taggers da biblioteca NLTK utilizam este corpus para treino. A tabela 4.1 mostra as tags e respetiva descrição utilizadas neste corpus. A utilização do léxico/dicionário de sentimentos para marcação das palavras com tag *None* permitiu marcar adicionalmente mais de 1000 palavras distintas do conjunto de dados em comparação com a utilização de apenas o modelo treinado (antes da aplicação da tag padrão).

Tag	Descrição	Tag	Descrição
n	nome/substantivo	pron-pers	pronome pessoal
prop	nome próprio	pron-det	pronome determinativo
adj	adjetivo	pron-indp	pronome independente
prp	preposição	adv	advérbio
v-fin	verbo finito	num	numeral
v-inf	verbo infinito	in	interjeição
v-pcp	verbo particípio	conj-s	conjunção subordinativa
v-ger	verbo gerúndio	conj-c	conjunção coordenativa
art	artigo		

Tabela 4.1: *Floresta Sintática POS tags.*

5 Modelos Supervisionados

As soluções baseadas em aprendizagem automática, supervisionada como não supervisionada, são utilizadas em todo o tipo de sistemas. A utilização deste tipo de soluções em análise de sentimentos tornou-se o standard, dadas as vantagens que possui em comparação com modelos mais tradicionais baseados em dicionários ou regras gramaticais manuais. Assim, desenvolvemos um conjunto de modelos de aprendizagem automática supervisionados de classificação. Cada modelo foi desenvolvido de modo a tentar melhorar o desempenho do modelo anterior, pela utilização de técnicas de aprendizagem automática adequadas ao processamento de linguagem natural e soluções comuns de análise de sentimentos em textos.

5.1 Técnicas de Modelação e Modelos Desenvolvidos

No desenvolvimento de uma qualquer solução baseada em aprendizagem automática é necessário utilizar vários classificadores de modo a testar qual deles se adequa melhor ao problema, pela qualidade do resultado das métricas definidas para avaliação dos modelos. Para além da utilização de vários classificadores, vários modelos podem também ser implementados, com variações nas transformações aplicadas aos dados e técnicas de implementação e à representação dos dados distintas entre eles. Apesar de extensivo este processo de implementação de vários modelos com vários classificadores, garante que se consegue um melhor modelo de entre um conjunto de possíveis modelos ou implementações. Foi neste sentido que foram desenvolvidos os modelos de aprendizagem automática supervisionados. Na concretização destes modelos implementámos um conjunto de 9 propostas com diferentes métodos de representação e de processamento dos comentários. Além disso, para cada um dos modelos desenvolvidos, foram utilizados 4 classificadores distintos (tabela 5.1).

Modelos	Classificadores
Presença de Palavras	
Frequência de Palavras	
TF-IDF	Naive Bayes
Frequência de Palavras (sem acentos e sufixos)	
Frequência de Palavras (sem acentos e sufixos e 2-gramas)	Support Vector Machines
Frequência de Palavras (sem acentos e sufixos e 3-gramas)	
Frequência de Palavras com Reamostragem dos Dados de Treino (sem acentos e sufixos e 2-gramas)	Random Forests
Word2Vec (sem acentos e sufixos)	XGBoost
Frequência de Palavras e POS Tags (sem acentos e sufixos)	

Tabela 5.1: Modelos desenvolvidos e classificadores utilizados.

Obteve-se, assim, 36 variações nos modelos desenvolvidos (9*4). Apesar deste número, aqui apenas vai ser apresentada informação detalhada sobre a implementação e resultados de alguns dos modelos, considerados mais relevantes, quer em termos da forma de implementação como resultados.

5.2 Condições de Treino e de Teste e Avaliação dos Modelos

Os modelos supervisionados de aprendizagem automática necessitam de dados para treino e para teste. Para isso, foi utilizado um conjunto de dados para este efeito, utilizando a técnica *train-test split*, com divisão aleatória e estratificada do mesmo: 70% dos comentários foram utilizados para treino e 30% para teste da qualidade dos modelos. Todos os modelos desenvolvidos utilizaram os mesmos dados de treino e de teste. A divisão estratificada garante que a mesma percentagem de comentários positivos e negativos é utilizada como a que ocorre no conjunto de dados, garantindo-se, assim, que não ocorrem casos extremos como os dados de treino serem quase todos totalmente constituídos pela classe positiva e os dados de teste pela classe negativa. Relembra-se que a distribuição das polaridades dos comentários do conjunto de dados é igual a 86% para os positivos e 14% para os negativos, sendo estas as distribuições tanto dos dados utilizados para treino como para teste.

Várias métricas podem ser utilizadas para testar a qualidade dos resultados de um modelo supervisionado. As mais comuns são *Accuracy* e *Precision* mas outras como *Sensitivity*, *Specificity*, *F1 Score* são também usadas de acordo com as características do conjunto de dados e dos dados de treino e teste dos modelos.

A escolha acerca de qual a métrica utilizar para avaliar os modelos não foi simples, devido ao não-balanceamento do conjunto de dados, já apresentado durante o pré-processamento, dado que leva a que muitas das métricas mais utilizadas não sejam possíveis de aplicar e confiar nos resultados de forma realista. Considere-se o exemplo abaixo da consequência da utilização de um conjunto de dados não-balanceado para treino de um modelo.

Exemplo 1: Considere-se um possível conjunto de dados binário não-balanceado, no qual uma classe tem 900 entradas e a outra 100.

Considere-se agora que um qualquer modelo prevê sempre a classe de maior frequência e de acordo com a matriz de confusão:

True Positive (**TP**) **900**	False Negative (**FN**) **0**
False Positive (**FP**) **100**	True Negative (**TN**) **0**

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{900}{100} = 0.9 = 90\%$$

Embora um valor de *accuracy* igual a 90% seja muito bom, na realidade é enganoso. O modelo não tem qualquer capacidade de previsão dado que prevê sempre a classe mais positiva. Isto acontece devido ao não-balanceamento dos dados pelas classes.

O que foi demonstrado no exemplo acima ocorre em muitas das métricas mais comuns. Neste sentido foi escolhida uma métrica capaz de lidar com classes não-balanceadas: *AUCROC* (*área under ROC curve*). Uma motivação para utilização desta métrica passou também pela consideração que ambas as classes (positiva e negativa) possuem o mesmo peso e importância ou seja, a consequência do modelo errar na previsão de um comentário positivo tem o mesmo peso como errar num comentário negativo.

5.3 Modelos & Resultados

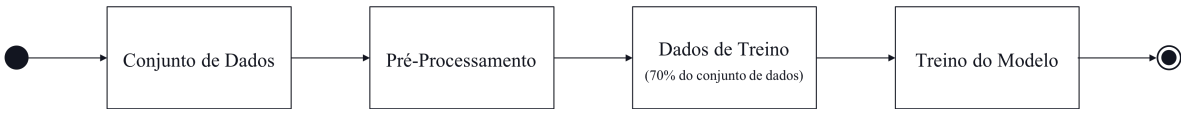


Figura 5.1: Esquema geral da construção dos modelos.

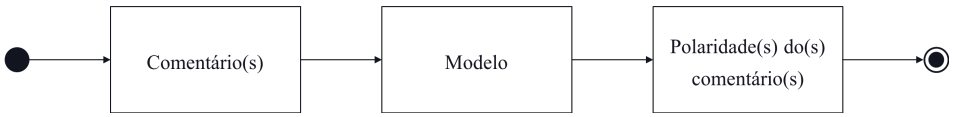


Figura 5.2: Esquema geral da utilização dos modelos.

Cada um dos modelos desenvolvidos segue um esquema geral de implementação e utilização idêntico. Antes do treino do modelo é aplicado o pré-processamento a todo o conjunto de dados (apresentado no capítulo 3) de modo a normalizar os dados para serem utilizados. Após isto, o conjunto de dados já normalizado é dividido de forma aleatória e estratificada para treino (70%) e teste (30%) dos modelos, como foi visto anteriormente. Dado que são modelos supervisionados, os dados de treino e de teste têm de estar anotados com os valores do atributo a prever. Naturalmente, este atributo representa a polaridade de cada comentário que tanto pode ser 0, que representa um comentário positivo, ou 1, que representa um comentário negativo. O resultado dos modelos é, portanto, binário e os pesos atribuídos a cada classe são iguais – superior a 50% é considerado um comentário positivo e inferior é negativo. O modelo é treinado com os dados de treino e pode depois ser utilizado para prever a polaridade de novos dados ou comentários. Os vários modelos desenvolvidos diferem, portanto, entre si na forma como os comentários são representados e processados no modelo.

5.3.1 Presença de Palavras e Frequência de Palavras

Um modelo de aprendizagem automática não consegue lidar de forma natural com texto, sendo necessário utilizar uma representação numérica do mesmo. Os comentários devem, então, ser transformados para este formato. No final do pré-processamento os comentários estão representados por listas de *tokens* que facilitam a construção de um vocabulário de todas as palavras distintas dos comentários.

A transformação dos comentários para uma representação numérica é efetuada pela sua vetorização, utilizando o vocabulário das palavras distintas. Portanto, constroi-se uma matriz, na qual cada coluna é uma palavra do vocabulário e cada linha é o vetor representativo de um comentário. O tamanho de cada vetor dos comentários é igual ao tamanho do vocabulário.

Considere-se o exemplo abaixo representativo deste processo.

Exemplo 1: Considere-se os comentários do conjunto de dados, já pré-processados, cujo estado final é uma lista de *tokens*:

C1: ['relação', 'preço', 'qualidade', 'muito', 'boa']

C2: ['muito', 'bom', 'ultrapassa', 'muito', 'expectativas']

C3: ['excelente', 'qualidade']

C4: ['bastante', 'satisfeita', 'relação', 'preço', 'qualidade']

C5: ['muito', 'bom', 'ecrã', 'excelente']

O vocabulário deste conjunto de comentários é constituído pelas palavras:

relação / preço / qualidade / muito / boa / bom / ultrapassa / expectativas / excelente / bastante / satisfeita / ecrã

A quantidade total de palavras de todos os comentários é igual a 21 e o tamanho do vocabulário é igual a 12. Existem, portanto, 9 palavras comuns entre os comentários.

A representação dos comentários em vetores, numa matriz, é dada por:

relação	preço	qualidade	muito	boa	bom	ultrapassa	expectativas	excelente	bastante	satisfeita	ecrã
1	1	1	1	1	0	0	0	0	0	0	0
0	0	0	1	0	1	1	1	0	0	0	0
0	0	1	0	0	0	0	0	1	0	0	0
1	1	1	0	0	0	0	0	0	1	1	0
0	0	0	1	0	1	0	0	1	0	0	1

Cada comentário é representado por um vetor no espaço do vocabulário.

A utilização da matriz de vetores permite que os modelos supervisionados utilizem cada umas das palavras do vocabulário como novos atributos. Quanto mais reduzido for um vocabulário (até um certo ponto) e representativo do problema de análise mais eficientemente, o modelo vai conseguir determinar padrões de convergência para as classes resultado (positivo ou negativo). Torna-se, portanto, evidente a importância da limpeza efetuada efetuada no conjunto de dados durante o pré-processamento, reduzindo o vocabulário pela remoção de palavras não características da presença de sentimentos. Este tipo de representação é comummente designada por *bag of words* e pode ser construída representando os vetores pela presença ou frequência das palavras. No primeiro caso, os valores possíveis dos vetores são 0 e 1, enquanto

que no segundo caso são 0 e o número de vezes que uma mesma palavra ocorre no comentário. A matriz apresentada no exemplo anterior tem uma representação em presença de palavras e a mesma matriz em frequência de palavras ficaria:

relação	preço	qualidade	muito	boa	bom	ultrapassa	expectativas	excelente	bastante	satisfeita	ecrã
1	1	1	1	1	0	0	0	0	0	0	0
0	0	0	**2**	0	1	1	1	0	0	0	0
0	0	1	0	0	0	0	0	1	0	0	0
1	1	1	0	0	0	0	0	0	1	1	0
0	0	0	1	0	1	0	0	1	0	0	1

Resultados

Os modelos foram treinados com 4 classificadores distintos (já apresentados na tabela 5.1). Os parâmetros de qualquer classificador são obtidos automaticamente pelo uso de uma *GridSearchCV* garantindo que independentemente do modelo e da representação dos comentários, os melhores parâmetros são sempre conseguidos. Uma *GridSearchCV* utiliza um conjunto de parâmetros definidos para cada classificador e avalia o modelo com todas as combinações possíveis dos parâmetros que, naturalmente, leva à necessidade de uma grande quantidade de tempo de treino. A obtenção automática dos melhores parâmetros é efetuada em todos os modelos desenvolvidos.

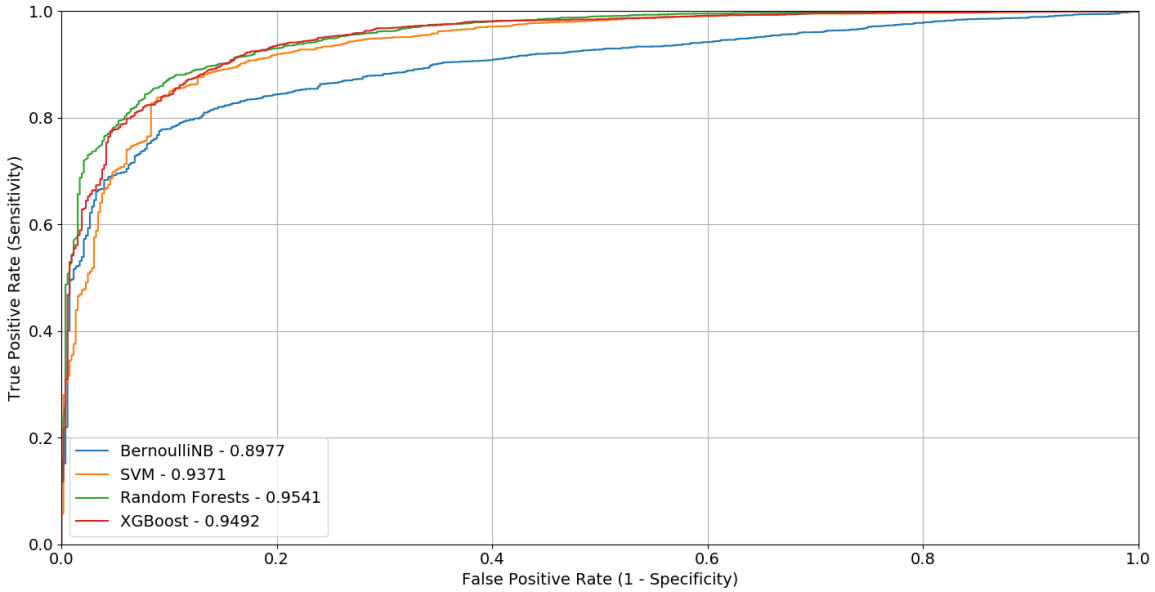


Figura 5.3: Modelo presença de palavras – curvas ROC.

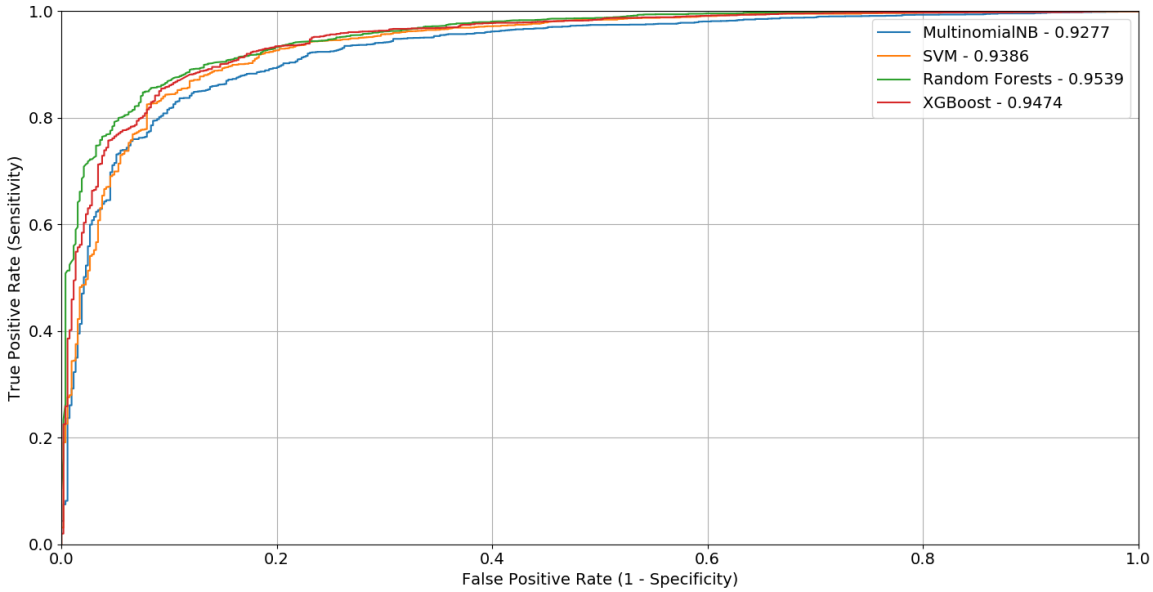


Figura 5.4: Modelo frequência de palavras – curvas ROC.

A avaliação dos modelos é efetuada com os dados de teste e determinada pelo melhor valor da AUCROC. As figuras 5.3 e 5.4 mostram os valores obtidos nos dois modelos para cada classificador. Em ambos os modelos, o classificador baseado em *Random Forests* foi o que conseguiu melhores resultados na AUCROC. Baseando-se nesta métrica e neste classificador, pode-se concluir que os modelos são idênticos na capacidade de previsão da polaridade de novos comentários (0.9541 vs 0.9539).

5.3.2 Frequência de Palavras sem Acentos e sem Sufixos

Como foi visto anteriormente, a qualidade de escrita dos comentários não é a melhor dada a sua origem suscetibilidade à ocorrência de erros ortográficos. No pré-processamento foi aplicado um conjunto de transformações para corrigir e minimizar a ocorrência e efeitos negativos destes casos mas, apesar disto, nem todos conseguem ser tratados. Dado que a língua portuguesa é caracteristicamente acentuada, muitos dos erros ortográficos ocorrem devido à ausência ou incorreta acentuação das palavras. Neste sentido, foram removidos os acentos de todas as palavras, eliminando qualquer erro ortográfico deste tipo bem como a redução do vocabulário pela consequente uniformização das palavras. As palavras estão sujeitas a flexões gramaticais levando a que uma palavra raiz possa ser escrita de várias formas. A utilização de sufixos transforma uma palavras raiz em várias formas e intensificam ou diminuem o seu significado. O inverso é também possível, a remoção de um sufixo a uma palavra transforma a mesma para o seu formato raiz mas o significado mantém-se imutável embora possa haver perda de

intensidade. A remoção de sufixos é, então, uma possível transformação capaz de uniformizar as palavras. A remoção dos sufixos das palavras da tabela abaixo mostra o resultado da aplicação do *SnowballStemmer* [2], utilizado para remover os sufixos de todas as palavras.

Com sufixos	Sem sufixos
felizmente	feliz
surpreendente	surpreendent
surpreendentemente	
descontente	descontent
descontentamento	
significativo	signific
significativamente	

Tabela 5.2: Remoção de sufixos – exemplos.

Tanto a remoção dos acentos como dos sufixos permitiu reduzir o vocabulário de acordo com a figura 5.5. A redução foi significativa, cerca de 43% em relação ao vocabulário original. Esta redução é também, refletida no tamanho dos vetores representativos dos comentários. O novo vocabulário e comentários são transformados numa matriz de vetores pelo mesmo método utilizado no modelo de frequência de palavras original.

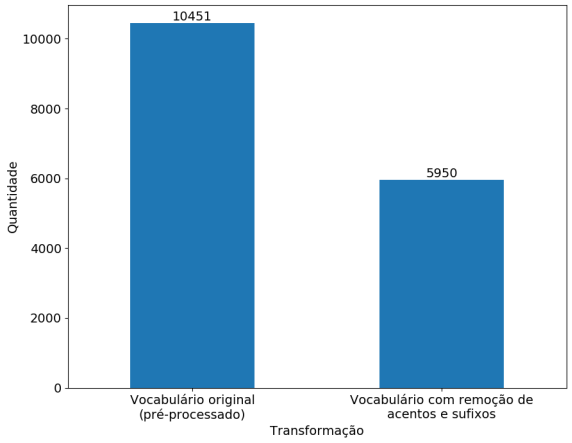


Figura 5.5: Tamanho do vocabulário após remoção de acentos e sufixos.

Resultados

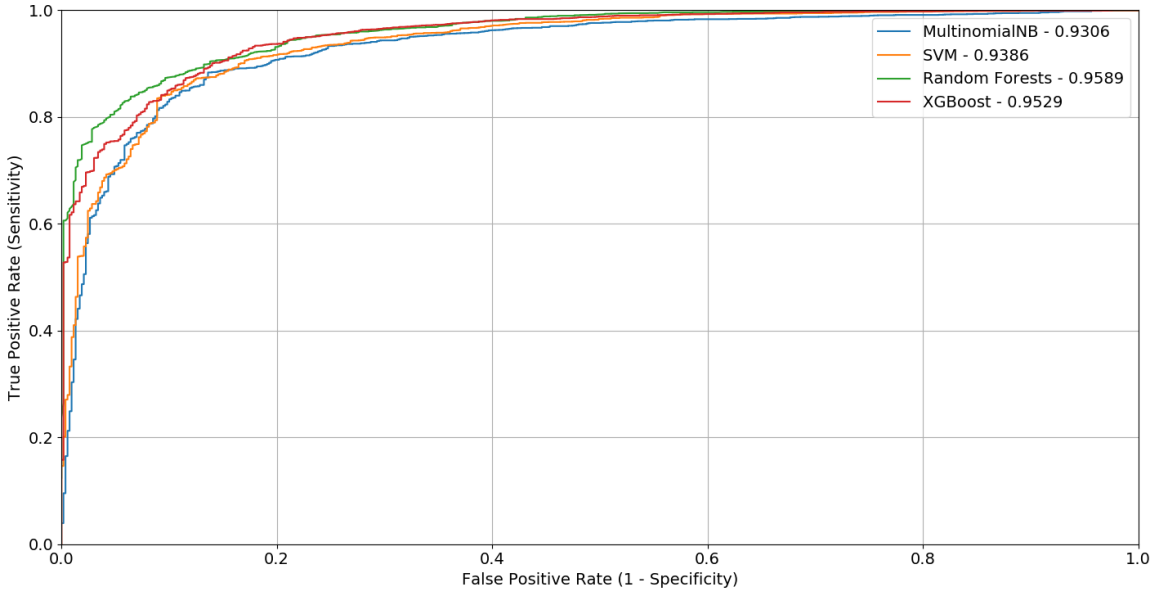


Figura 5.6: Modelo frequência de palavras sem acentos e sem sufixos – curvas ROC.

Tal como nos modelos anteriores, *Random Forests* foi o classificador que conseguiu melhores resultados, seguido por *XGBoost*, *SVM* e por fim *Naive Bayes*.

5.3.3 Frequência de Palavras sem Acentos e sem Sufixos, 2-gramas e Reamostragem dos Dados de Treino

Dando continuação ao modelo anterior apresentado, no presente modelo foram efetuadas mais duas transformações:

- Representação dos comentários em conjuntos de 2-gramas
- Reamostragem dos dados de treino

A representação em conjuntos de n-gramas, neste caso 2, permite que seja preservada alguma da estrutura dos comentários como a posição das palavras. Isto pode ser especialmente útil para os comentários negativos. Como foi visto no pré-processamento, muitos dos comentários negativos são formados pela negação de palavras de sentimento positivas, a utilização de 2-gramas pode aumentar a capacidade de os modelos identificarem estes casos, dado que uma palavra de negação está frequentemente junta à palavra que afeta, podendo esta relação ser facilmente captada por conjuntos de 2-gramas. O exemplo abaixo mostra a transformação em bigramas dos comentários já apresentados no exemplo 1 da secção 5.3.1 e a matriz de

frequência de palavras desses bigramas. Note-se que antes da transformação para bigramas foi efetuada a remoção de acentos e sufixos às palavras mas no exemplo tal não acontece para facilidade de leitura/análise.

Exemplo 2: Considere-se a representação em bigramas dos comentários anteriormente apresentados:

C1: ['relação preço', 'preço qualidade', 'qualidade muito', 'muito boa']

C2: ['muito bom', 'bom ultrapassa', 'ultrapassa muito', 'muito expectativas']

C3: ['excelente qualidade']

C4: ['bastante satisfeita', 'satisfeita relação', 'relação preço', 'preço qualidade']

C5: ['muito bom', 'bom ecrã', 'ecrã excelente']

O vocabulário destes comentários representados em bigramas é constituído por:

relação preço / preço qualidade / qualidade muito / muito boa / muito bom / bom ultrapassa / ultrapassa muito / muito expectativas / excelente qualidade / bastante satisfeita / satisfeita relação / bom ecrã / ecrã excelente

A quantidade total de bigramas de todos os comentários é igual a 16 e o tamanho do vocabulário é igual a 13. Existem, portanto, 3 bigramas comuns entre os comentários.

Os comentários são representados em vetores numa matriz de frequência de palavras/bigramas dada por:

relação preço	preço qualidade	qualidade muito	muito boa	muito bom	bom ultrapassa	ultrapassa muito	muito expectativas	excelente qualidade	bastante satisfeita	satisfeita relação	bom ecrã	ecrã excelente
1	1	1	1	0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0
1	1	0	0	0	0	0	0	0	1	1	0	0
0	0	0	0	1	0	0	0	0	0	0	1	1

A reamostragem dos dados tem como objetivo eliminar o impacto negativo do não balanceamento das classes do conjunto de dados. A elevada quantidade de comentários positivos em comparação com os negativos pode afetar a capacidade de aprendizagem/treino dos modelos fazendo com que os mesmos sejam tendenciosos para a classe mais frequente.

Várias formas de reamostragem dos dados podem ser aplicadas. A utilização de uma percentagem dos dados da classe maioritária igual à percentagem da classe minoritária é uma solução

redutiva na qual são excluídos dados. Apesar de simples, esta solução remove dados possivelmente importantes da classe maioritária que poderiam ser relevantes para o treino e teste do modelo. A solução utilizada segue o inverso desta. Nesta solução foi utilizado *SMOTE*, que é uma técnica de reamostragem de aumento da classe minoritária para a mesma percentagem da maioritária. Esta técnica constrói um conjunto de novas entradas com ligeiras diferenças em relação aos comentários do conjunto de dados. Estas novas entradas são construídas a partir dos comentários da classe minoritária que, em conjunto com a mesma, iguala a quantodade de comentários da classe maioritária. Duas possibilidades foram consideradas na forma de aplicação de *SMOTE* sobre os dados:

1. Aplicação sobre todo o conjunto de dados e posterior divisão em treino e teste.

2. Divisão dos dados em treino e teste e aplicação de *SMOTE* apenas sobre os dados de treino.

A segunda opção foi a escolhida dado ser a mais robusta. Ao aplicar o método de reamostragem apenas aos dados de treino garante-se que o modelo é treinado com dados balanceados e os dados de teste não são afetados pelo processo de reamostragem, mantendo as características originais dos comentários. O modelo é, portanto, garantidamente testado com dados originais do conjunto de dados sem ocorrência de dados artificiais (resultantes da aplicação de *SMOTE*) que aconteceria na primeira opção.

Resultados

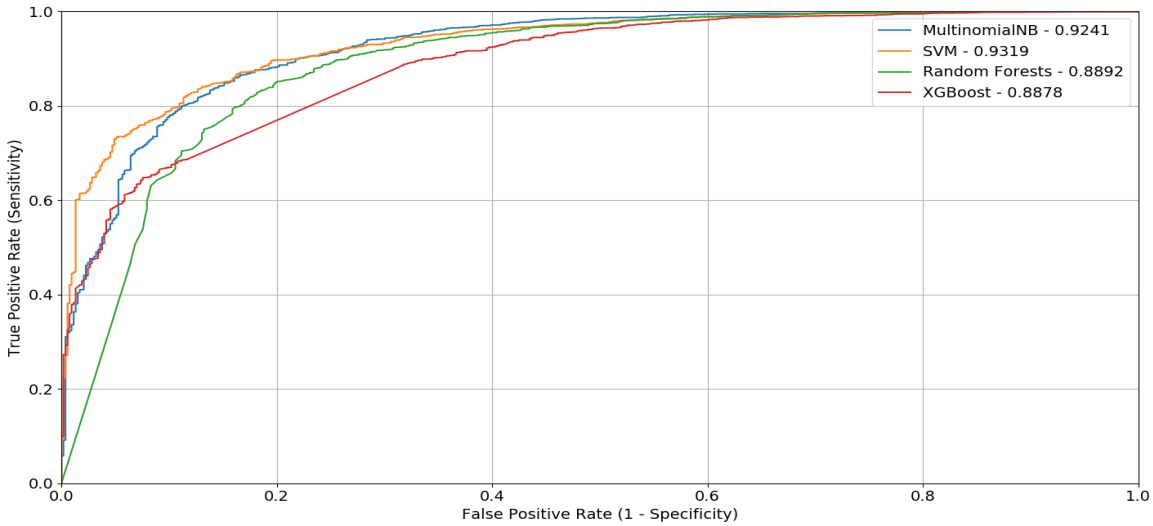


Figura 5.7: Modelo frequência de palavras sem acentos e sem sufixos, 2-gramas e reamostragem dos dados de treino – curvas ROC.

SVM foi, agora, o melhor classificador, seguido por *Naive Bayes*, *Random Forests* e *XGBoost*. As transformações adicionais implementadas neste modelo (bigramas e reamostragem dos dados de treino) não conseguiram melhorar os resultados de quaisquer classificador em comparação com o modelo inicial apenas com remoção de acentos e sufixos. *Random Forests* e *XGBoost* tiveram até uma redução muito acentuada.

5.3.4 Word2Vec (sem acentos e sem sufixos)

Como tem sido mostrado nos modelos anteriores, os vetores utilizados para representar os comentários são utilizados pelos modelos para treino e teste. Quanto mais representativo for um vetor do respetivo comentário mais facilmente um modelos consegue aprender. Para além disto, é importante que os vetores dos comentários de uma mesma classe sejam similares para que um modelo consiga mais facilmente convergir os comentários para a classe correta. O Word2Vec foi aqui utilizado de forma a tentar conseguir isto mais facilmente.

Word2Vec é um modelo baseado em redes neuronais desenvolvido pela *Google* que permite representar cada palavra como um vetor que dado um contesto onde a palavra se encontra (conjunto de outras palavras como comentários) consegue determinar um conjunto de palavras similares de entre um conjunto de dados. Isto pode ser facilmente aplicado sobre comentários (constituídos por conjuntos de palavras) em vez de palavras individuais. Word2Vec é então uma solução de transformação de texto em vetores especialmente determinados tendo em vista a similaridade das palavras dentro de um contexto.

Existe um conjunto de modelos já treinados em Word2Vec para um grande espectro de problemas, incluindo a representação de sentimentos, com conjuntos de dados otimizados para o problema. Contudo, novamente, este tipo de recursos é muito limitado ou escasso quando a linguagem alvo não é o inglês. Isto levou a que fosso desenvolvido todo o modelo de raiz, desde o seu treino do mesmo até à representação dos comentários em vetores. Uma das vantagens de utilização de Word2Vec é o facto de não necessitar de dados de treino anotados. Os vetores representativos de cada palavra são obtidos de acordo com o contexto no qual a palavra se insere, fazendo com que seja necessário apenas texto formatado para treino do modelo. A figura 5.8 mostra a estrutura do modelo desenvolvido.

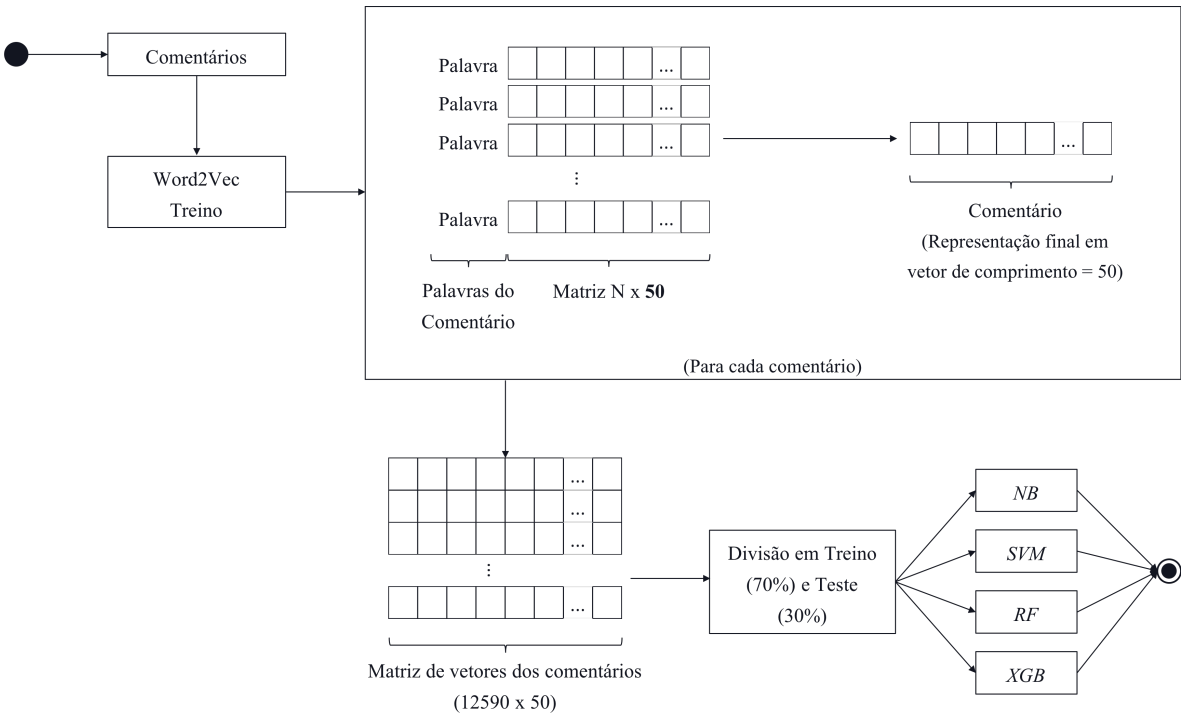


Figura 5.8: Esquema do modelo baseado em Word2Vec.

O modelo Word2Vec é treinado com todos os comentários do conjunto de dados, para se obter um vetor para cada comentário, representativo do mesmo e baseado na similaridade entre as palavras. Após o treino do modelo Word2Vec, cada palavra de cada comentário é representada por um vetor de tamanho igual a 50 (valor escolhido como parâmetro) e cada comentário é, então, representado pelo conjunto das palavras que o constituem e representado por uma matriz de vetores com tamanho *Nx50*, onde *N* é o número de palavras do comentário. Esta matriz é então convertida num único vetor, utilizando a média dos valores da matriz, representativo do comentário e de tamanho igual a 50. Consegue-se desta forma representar um comentário num vetor único e representativo da similaridade entre comentários dado que é o resultado dos vetores das palavras individuais que o constituem, vetores estes determinados pela similaridade entre as palavras. Os vetores finais de cada comentário são representados numa matriz com dimensões *12590x50* (quantidade total de comentários do conjunto de dados por tamanho definido para os vetores), a matriz é dividida em dados de treino e teste de igual forma como nos modelos anteriormente desenvolvidos e são depois treinados os modelos de aprendizagem automática. Note-se a distinção entre a matriz de vetores dos comentários aqui utilizada e a matriz de frequência de palavras utilizada em modelos anteriores. Ao contrário dos modelos anteriores, a matriz aqui utilizada possui um número de colunas reduzido (50) em vez de o tamanho do vocabulário, representando a similaridade entre os comentários e não a frequência com que uma palavra ocorre num comentário.

Resultados

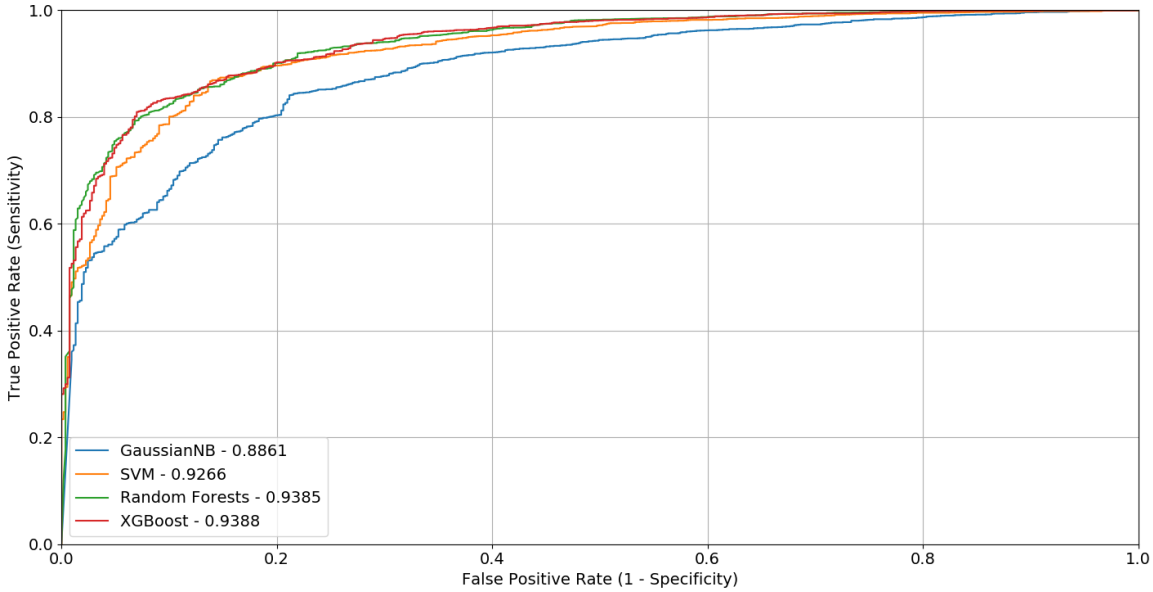


Figura 5.9: Modelo Word2Vec (sem acentos e sem sufixos) – curvas ROC.

Random Forests e *XGBoost* foram os modelos com melhores resultados e podem ser considerados idênticos dada a grande proximidade dos mesmos. Apesar da utilização de Word2Vec para representação dos comentários em vetores de similaridade, os resultados obtidos pelos classificadores seguem um intervalo de valores idêntico aos resultados de modelos anteriores. Um possível bottleneck capaz de explicar a falta de melhores resultados é provocado pelos dados utilizados para treino do modelo Word2Vec. Como referido anteriormente, o Word2Vec é um modelo baseado em redes neuronais e como qualquer modelo baseado nesta técnica, são necessárias grandes quantidades de dados de treino para que o modelo consiga distinguir e convergir os dados de forma mais eficaz que os tradicionais modelos de aprendizagem automática não baseados em redes neuronais. O tamanho do conjunto de dados utilizado (12590 comentários) é reduzido para utilização numa solução baseada em redes neuronais e, o facto do mesmo ser não balanceado, faz com que reduza ainda mais a capacidade de distinção entre os comentários.

5.3.5 Frequência de Palavras e POS Tags sem Acentos e sem Sufixos

Este modelo segue uma estrutura semelhante aos modelos anteriores (exceto ao do Word2Vec). A fase de pré-processamento permitiu não só tratar os dados como também excluir uma grande parte deles, não relevantes para análise de sentimentos, diminuindo o ruído e o tamanho do vocabulário. O mesmo acontece com a remoção de acentos e sufixos. A utilização de POS tags

segue também este objetivo de diminuir o vocabulário pela exclusão de palavras não relevantes para o tema de análise.

Para marcar as palavras de todos os comentários foi utilizado o *POS Tagger* desenvolvido, após o pré-processamento, com as respetivas partes do discurso. Uma vez marcados os comentários, foram removidas todas as palavras que não fossem:

- Adjetivos
- Verbos (todos os tipos)
- Advérbios
- Nomes (apenas se nenhuma das anteriores)

São mantidas apenas as palavras com estas partes do discurso dado que são as mais representativas da presença de sentimentos/opiniões, como já foi discutido no capítulo 2. Os *nomes* são apenas utilizados caso um comentário não possua qualquer palavra com as restantes partes do discurso, impedindo assim que um comentário fique vazio (relembra-se que um comentário é representado numa lista de *tokens*/palavras). O exemplo abaixo mostra a aplicação do *POS Tagger* desenvolvido sobre os comentários já apresentados no exemplo 1 da secção 5.3.1 e a respetiva matriz de frequência de palavras. Note-se que a remoção de acentos e sufixos é aplicada depois do *POS Tagger* mas não é refletida no exemplo para facilidade de leitura/análise.

Exemplo 3: Considere-se o resultado da aplicação do *POS Tagger* sobre os comentários anteriormente apresentados e posterior remoção das palavras que não sejam nenhuma das partes do discurso atrás apresentadas:

C1: ['muito', 'boa']

C2: ['muito', 'bom', 'ultrapassa', 'muito']

C3: ['excelente']

C4: ['bastante', 'satisfeita']

C5: ['muito', 'bom', 'excelente']

O vocabulário é constituído pelas palavras:

muito / boa / bom / ultrapassa / excelente / bastante / satisfeita

A quantidade total de palavras dos comentários é igual a 12 e o tamanho do vocabulário é igual a 7. Existem, portanto, 5 palavras comuns entre os comen-

tários.

A representação dos comentários em vetores, numa matriz, é dada por:

muito	boa	bom	ultrapassa	excelente	bastante	satisfeita
1	1	0	0	0	0	0
2	0	1	1	0	0	0
0	0	0	0	1	0	0
0	0	0	0	0	1	1
1	0	1	0	1	0	0

Cada comentário é representado por um vetor no espaço do vocabulário.

Resultados

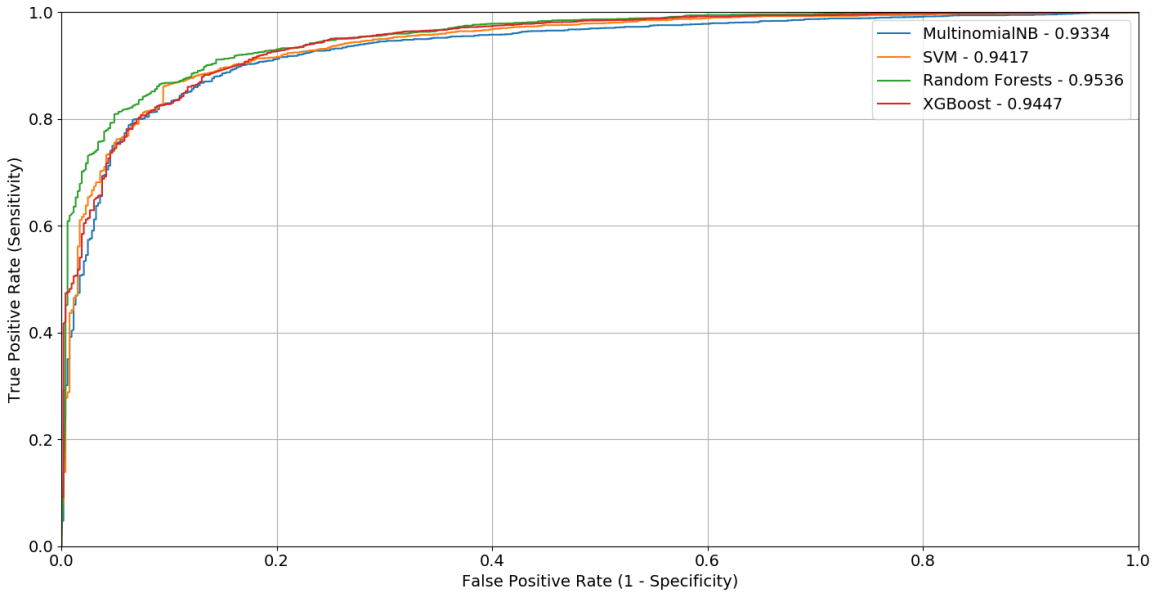


Figura 5.10: Modelo Frequência de Palavras e POS Tags sem Acentos e sem Sufixos – curvas ROC.

Os resultados seguem um padrão semelhante à mairia dos resultados dos restantes modelos, nos quais *Random Forests* é o melhor classificador, seguido por *XGBoost* e *SVM*. *Naive Bayes* é o classificador com menores resultados.

5.4 Análise dos Resultados e o Melhor Modelo

Nesta secção procura-se analisar os resultados como um todo, pela observação da distribuição geral e pela comparação entre os valores obtidos nos vários classificadores de cada modelo. Esta análise é indispensável para conseguir definir um melhor modelo final de aprendizagem automática supervisionada. A figura 5.11 mostra a distribuição do valor da AUCROC, métrica utilizada para avaliar os modelos.

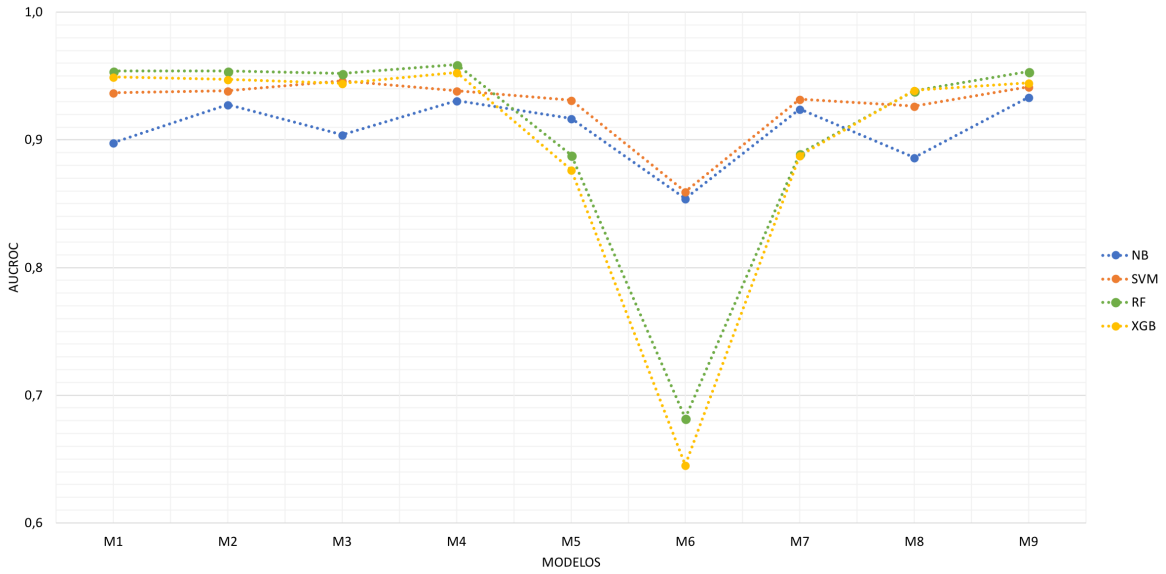


Figura 5.11: Distribuição dos valores de AUCROC obtidos pelos classificadores nos modelos.

Na figura acima é necessária especial atenção para o valor de início do eixo dos valores da AUCROC. O intervalo de valores possíveis para esta métrica está entre 0 e 1, tendo sido escolhido um valor inicial igual a 0.6, dado que nenhum modelo obteve um resultado inferior a este, o que facilita a análise do gráfico. O eixo dos modelos representa cada um dos modelos desenvolvidos e de acordo com a correspondência da tabela abaixo.

M1	Presença de Palavras
M2	Frequência de Palavras
M3	TF-IDF
M4	Frequência de Palavras (sem acentos e sem sufixos)
M5	Frequência de Palavras (sem acentos e sem sufixos e 2-gramas)
M6	Frequência de Palavras (sem acentos e sem sufixos e 3-gramas)
M7	Frequência de Palavras com Reamostragem dos Dados de Treino (sem acentos e sem sufixos e 2-gramas)
M8	Word2Vec (sem acentos e sem sufixos)
M9	Frequência de Palavras e POS Tags (sem acentos e sem sufixos)

A distribuição mostra que todos os modelos possuem valores de AUCROC elevados/bons para

qualquer classificador, nunca inferiores a aproximadamente 0.85 com exceção dos classificadores *RF* e *XGB* no modelo M6. Observa-se ainda que ao longo dos modelos, os resultados entre os classificadores seguem um padrão semelhante, com exceção de *SVM* em M3 e *RF*, *XGB* em M8, onde para cada modelo os valores de AUCROC ou aumentam ou diminuem para todos os classificadores. Aqui, existe uma diminuição significativa em todos os classificadores entre os modelos M5 e M7, inclusive. Estes três modelos têm uma característica em comum e que não ocorre nos restantes: conjuntos de n-gramas. Isto leva a concluir que esta transformação específica não deve ser utilizada para soluções de análise de sentimentos no contexto de análise utilizado (comentários de produtos).

À medida que os modelos foram desenvolvidos foram aplicadas novas transformações aos dados que são frequentemente utilizadas em soluções de análise de sentimentos em textos. Contudo, a distribuição dos resultados mostra que a utilização destas transformações, como *TF-IDF*, *n-gramas*, *Word2Vec*, *POS Tags*, produz resultados inferiores em comparação com o modelo anterior que faz menor ou nenhum uso destas transformações típicas. A tabela 5.3 mostra os valores da AUCROC obtidos em cada modelo por cada classificador. Aplica-se a correspondência atrás apresentada.

	M1	M2	M3	M4	M5	M6	M7	M8	M9
NB	0,8977	0,9277	0,9042	0,9306	0,9168	0,8540	0,9241	0,8861	0,9334
SVM	0,9371	0,9386	0,9464	0,9386	0,9313	0,8593	0,9319	0,9266	0,9417
RF	0,9541	0,9539	0,9522	0,9589	0,8883	0,6819	0,8892	0,9385	0,9536
XGB	0,9492	0,9474	0,9441	0,9529	0,8766	0,6453	0,8878	0,9388	0,9447

Tabela 5.3: Valores de AUCROC obtidos pelos classificadores nos modelos.

Os valores em destaque na tabela representam os mais elevados/melhores para cada classificador. *Random Forests* é o classificador com maior AUCROC seguido por *XGBoost*, *SVM* e por último *Naive Bayes*. Isto pode também ser observado na figura 5.11, onde o valor mais elevado de AUCROC ocorre em M4 com *Random Forests*. *Random Forests* é na verdade o melhor classificador para praticamente todos os modelos, como mostra a figura, *XGBoost* segue um padrão muito idêntico a *RF* com resultados ligeiramente inferiores. *Naive Bayes* e *SVM* têm também um padrão idêntico mas com diferenças maiores entre os resultados.

O modelo *Random Forests* foi, portanto, o classificador que obteve o maior valor de AUCROC que ocorre em M4 – Frequência de Palavras sem Acentos e sem Sufixos. Este modelo é então o melhor modelo em conjunto com *RF* para o contexto de análise. A figura 5.12 e a tabela 5.4 apresentam a matriz de confusão e informação adicional de outras métricas para além de AUCROC, respetivamente, sobre os dados de teste[1].

[1]Download dos resultados de todos os modelos disponível em: https://cl.ly/c440d8f14b8f

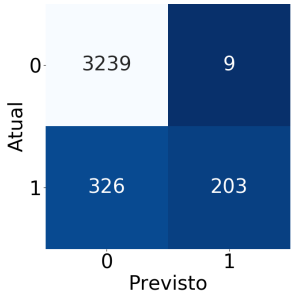


Figura 5.12: Matriz de Confusão (do melhor modelo).

Accuracy / Null_Accuracy	0.9113 / 0.8599
Precision	0.9086
Sensitivity	0.9972
Specificity	0.3837
AUCROC	0.9589

Tabela 5.4: *Accuracy*, *Precision*, *Sensitivity*, *Specificity* e *AUCROC* (do melhor modelo).

Considere-se na tabela 5.4 que *Null_Accuracy* é o valor de *Accuracy* para o modelo mais simples possível que classifica sempre os comentários com a classe mais frequente (positiva) e *Accuracy* é o valor obtido nesta métrica. Como foi demonstrado anteriormente, a utilização de *Accuracy* em conjuntos de dados não balanceados não é a melhor solução possível. Contudo, quando comparada com a *Null_Accuracy*, esta permite perceber/garantir se as previsões obtidas pelo modelo são significativamente melhores do que a previsão mais simples possível de classificar tudo com a classe mais frequente. Note-se ainda o valor elevado de *Null_Accuracy* que novamente comprova que esta métrica não é a mais adequada no tipo de conjunto de dados utilizado. Os valores obtidos nestas duas métricas mostram uma diferença significativa que permite garantir que as previsões são significativas, não aleatórias e melhores do que a previsão de todos os dados como positivos. Apesar disto, o modelo não é perfeito. A matriz de confusão mostra que a quantidade de falsos negativos (FN) é bastante reduzida (9). Contudo, a quantidade de falsos positivos (FP) é ainda relativamente elevada (326) que representa uma possível tendência do modelo para a classe positiva. Isto pode ser observado também na tabela 5.4 pelo valor da especificidade (métrica representativa da classe negativa) que é relativamente baixo principalmente quando comparado com o valor da sensibilidade (métrica representativa da classe positiva). Futuras melhorias ao modelo devem, então, focar-se em formas de representação dos comentários capazes de diferenciar melhor as duas polaridades dos comentários. Uma vez considerado como o melhor modelo, o mesmo foi guardado de forma persistente de modo a poder ser utilizado na previsão da polaridade de novos comentários e inserido em novos sistemas de análise de sentimentos.

5.5 Modelo de Classificação Não-Supervisionado e "Híbrido"

Embora, não muito frequente, a utilização de modelos não-supervisionados de aprendizagem automática sobre problemas cujo contexto seja baseado em texto, é também possível de implementar. A maior vantagem da utilização destes modelos é a não necessidade de conjuntos de dados anotados para treino. Em contrapartida, a utilização por si só não permite, na maior parte dos casos, classificar os dados em classes mas sim obter outro tipo de informação como a distribuição dos dados por semelhança entre eles (clustering). Assim, foi implementado um modelo clustering *K-Means* sobre o conjunto de dados, de acordo com a representação da figura 5.13, que mostra o processo de clustering e o sistema implementado.

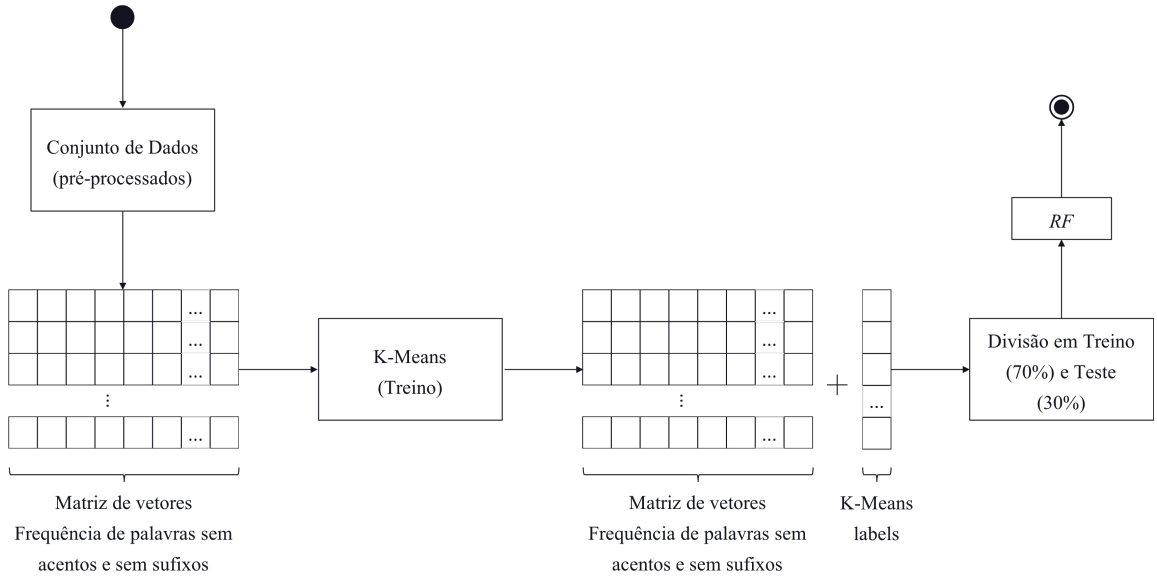


Figura 5.13: Estrutura modelo *K-Means* e modelo híbrido.

A aplicação de clustering em análise de sentimentos pode ser utilizada como forma de separação ou de distinção automática dos comentários por polaridade. As polaridades são, então, tratadas como clusters e os comentários automaticamente atribuídos a cada cluster de acordo com a similaridade entre eles. Tal como acontecia nos modelos supervisionados, o clustering não consegue processar diretamente o texto dos comentários, o que faz com que seja necessário utilizar uma representação numérica é necessária. Assim, foi utilizada a representação em matriz de vetores da mesma forma como nos modelos supervisionados, representativa da frequência de palavras sem acentos e sem sufixos, dado que esta representação foi a que obteve melhor AUCROC nos modelos supervisionados. Isto permitiu defender que é a melhor representação possível para distinção dos comentários por polaridade.

O algoritmo *K-Means* foi treinado com a matriz de vetores de frequência de palavras sem

acentos e sem sufixos. Um dos problemas na utilização deste tipo de clustering é a definição do número de clusters (valor K), que aqui foi facilmente ultrapassado, dado que o objetivo era o de separar os comentários em clusters de polaridade – foram utilizadas duas polaridades (positivo e negativo) –, logo o valor de K foi definido igual a 2. Uma vez treinado, dois resultados principais são obtidos: centroides e labels.

- **Centroides** – Ponto central de cada cluster formado entre o espaço n-dimensional utilizado para treino e associação de cada comentário a um cluster (n = nº de atributos = nº colunas da matriz de vetores = 5950).

- **Labels** – Vetor de tamanho igual ao nº de comentários com valores numéricos de 0 a K representativos do cluster a que cada comentário pertence.

Num modelo de clustering a forma mais simples de analisar os seus resultados do mesmo é pela representação gráfica dos clusters formados comummente num plano 2D. Como o número de atributos utilizados no treino do modelo é superior a 2 não é possível representar os comentários e os clusters diretamente num gráfico. Para contornar esse problema foi efetuada uma redução da dimensão da matriz de vetores para uma dimensão 2D que permite representar graficamente os comentários de acordo com as polaridades originais do conjunto de dados e os clusters obtidos com os comentários por eles distribuídos. As figuras 5.14 e 5.15 mostram esta informação, respetivamente (as cores utilizadas representam as duas polaridades/clusters).

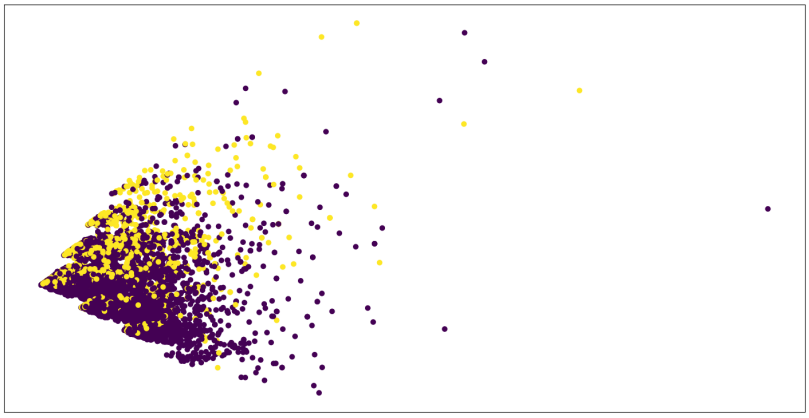


Figura 5.14: Distribuição dos comentários por polaridade.

A distribuição dos comentários por polaridade original (figura 5.14) mostra que não existem separações bem definidas entre as polaridades, possivelmente devido à semelhança entre os comentários positivos e negativos anteriormente discutida (muitos dos comentários negativos são formados pela negação de comentários positivos e não pela utilização de palavras de sentimento negativas). Isto pode influenciar negativamente a capacidade de separação dos comentários nos respetivos clusters em *K-Means*. A figura 5.15 mostra os clusters formados por *K-Means* e

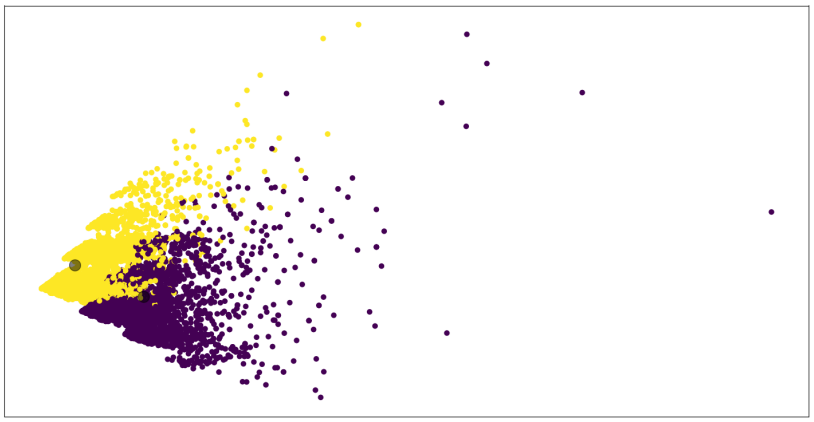


Figura 5.15: Clusters K-Means (K = 2).

os respetivos centroides. A distribuição ideal de *K-Means* seria igual à da figura 5.14. Apesar das distribuições das duas figuras não serem muito idênticas, *K-Means* conseguiu separar os comentários em dois clusters o mais aproximadamente possível à distribuição original (veja-se a correspondência entre as cores das distribuições).

Apesar dos resultados de *K-Means* não serem os melhores possíveis, estes podem ser utilizados para tentar melhorar o modelo de aprendizagem automática supervisionada anteriormente implementado. Foi, então, adicionado o vetor das labels de *K-Means* à matriz de vetores de frequência de palavras sem acentos e sem sufixos como um novo atributo. A nova matriz é dividida em dados de treino e teste iguais aos utilizados no modelo supervisionado, sendo treinado um classificador de *Random Forests*. Obersa-se, assim, um modelo supervisionado desenvolvido de igual forma como o melhor modelo supervisionado conseguido anteriormente, treinado sob o mesmo conjunto de dados de treino, mas agora com um atributo adicional que é o resultado do *K-Means* – modelo híbrido. A vantagem de utilização de *K-Means*, em vez de outro tipo de clustering, é a possibilidade de utilização para previsão e associação de novos dados aos clusters determinados durante o treino do modelo. Esta característica é essencial para a possível futura aplicação do modelo híbrido sobre novos dados.

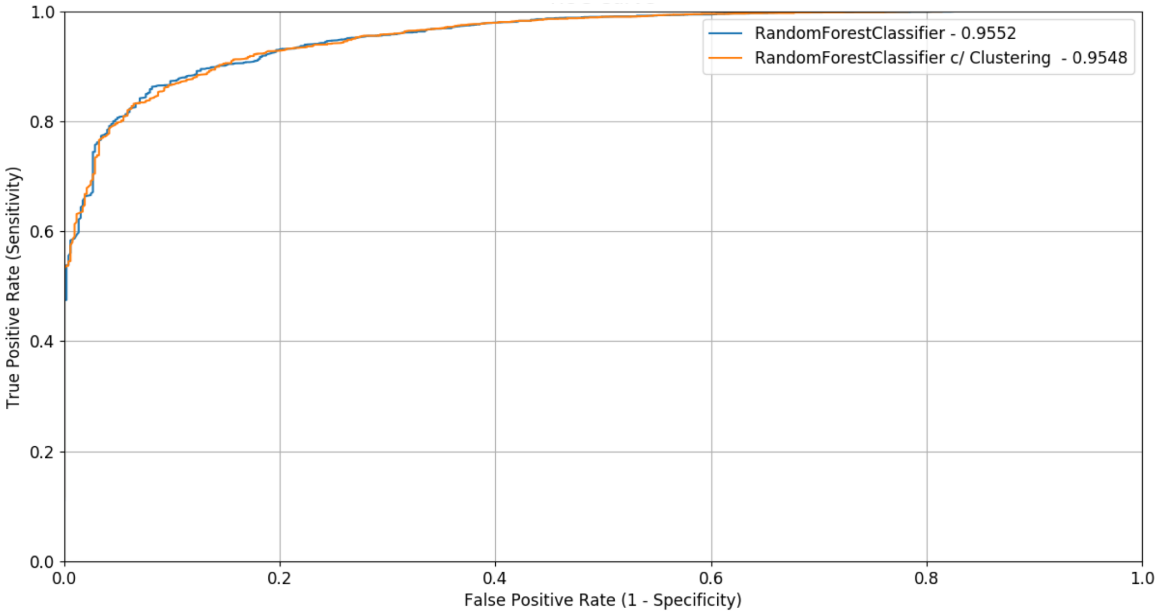


Figura 5.16: Melhor modelo supervisionado vs modelo com K-Means (híbrido) – curvas ROC.

A figura 5.16 mostra a comparação dos valores de AUCROC entre o melhor modelo supervisionado, conseguido anteriormente, e o novo modelo que faz uso de clustering *K-Means*. Os valores mostram que apesar da utilização dos resultados de *K-Means* como um novo atributo dos dados de treino, não se conseguiu melhorar o modelo anterior. A qualidade dos resultados de *K-Means* não é, portanto, suficientemente boa para poder ser utilizado. Dado estes resultados foram ainda efetuadas variações à implementação de *K-Means*, na qual foi efetuada a normalização dos dados e ainda a transformação dos vetores para a semelhança/distância cossena (*cosine similarity*), mas os resultados foram todos similares e não superiores ao modelo supervisionado original. Tentou-se ainda outro tipo de clustering – *SpectralClustering* – dado que este podia adaptar-se melhor à distribuição dos dados do que o *K-Means*. Contudo, o tempo de execução deste tipo de clustering é dado em O^3 com o número de atributos utilizados para treino e como este número é igual a 5950 o tempo necessário para o treino do modelo excedia um valor considerado razoável/adequado para a máquina utilizada no desenvolvimento do modelo.

6 Modelo Baseado em Dicionários

Os modelos de análise de sentimentos em textos baseados em aprendizagem automática supervisionada conseguem resultados com uma precisão bastante elevada, mas não são muito flexíveis na forma de implementação, sustentabilidade e diversidade do tipo de resultados que apresentam. Todavia, os modelos baseados em dicionários/léxicos de sentimentos são muito mais flexíveis, na forma de implementação, mais fáceis de manter e atualizar a longo prazo e permitem mais facilmente mostrar mais facilmente a informação extra, para além da simples polaridade binária de um comentário ou texto de análise, mas frequentemente com menor precisão dos resultados.

Um dos modelos mais simples baseado em dicionários de sentimentos utiliza a consulta direta ao dicionário para obter a polaridade das palavras reconhecidas. Contudo, outros fatores têm um grande impacto sobre a polaridade de certas palavras e, como tal, devem ser tidos em consideração para a implementação de um modelo o mais robusto possível. O modelo desenvolvido vai de encontro a esses fatores, fazendo uso de várias formas de tratamento de texto, obtenção de palavras representativas de sentimentos e identificação da polaridade de acordo com fatores próximos das palavras. A figura 6.1 mostra a arquitetura desse modelo.

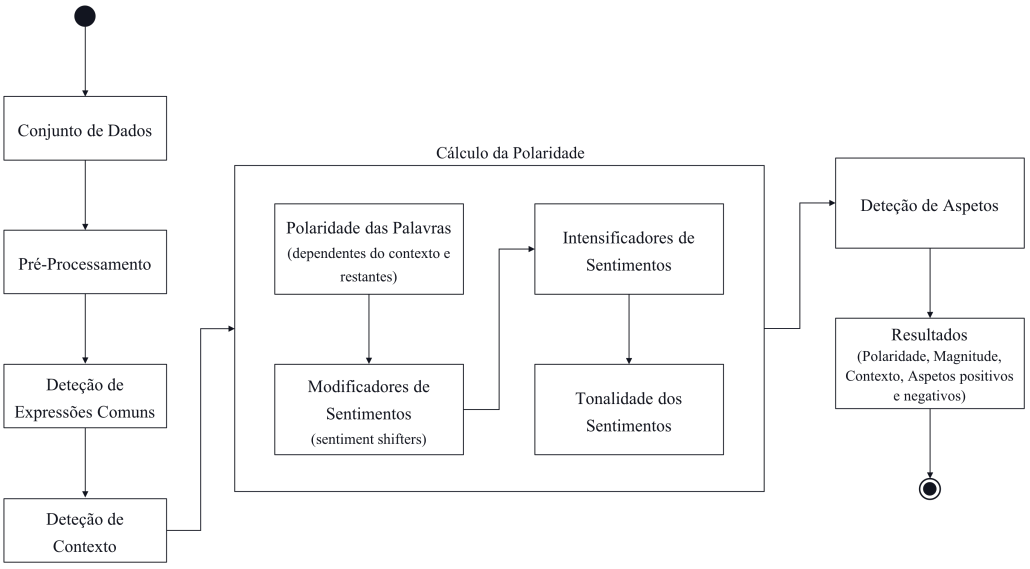


Figura 6.1: Arquitetura modelo baseado em dicionários.

O modelo foi desenvolvido ao nível do aspeto (maior nível possível segundo a literatura) e como tal possui também os níveis inferiores a este (documento e frase). Todas as transformações e métodos desenvolvidos necessitaram de uma grande revisão manual do dicionário de sentimentos e dos comentários do conjunto de dados, de modo a perceber a forma como os mesmos são escritos, e como são representadas as palavras de sentimento e a sua polaridade das mesmas consoante o contexto envolvente.

6.1 Abordagem e Técnicas de Processamento

Como em qualquer outro modelo, os comentários têm de ser pré-processados antes de serem utilizados. As mesmas transformações do pré-processamento aplicado nos modelos supervisionados são utilizadas aqui com a exceção de duas: a remoção de sinais de pontuação, caracteres especiais e números, e a remoção de palavras dependentes do tamanho. A remoção de *stop words* é ainda ligeiramente diferente e de acordo com a tabela 3.14 já apresentada anteriormente no capítulo 3. A não remoção dos dados destas transformações deve-se à natureza dos modelos baseados em dicionários e, por conseguinte, ao modelo desenvolvido. Ao contrário do que acontece com os modelos supervisionados, nos quais esta informação é considerada como ruído no processo de aprendizagem, os modelos baseados em dicionários fazem frequentemente uso da estrutura total dos textos onde, por exemplo, sinais de pontuação podem ser utilizados para separação de um documentos em frases individuais e a remoção de palavras por tamanho pode eliminar outras que potencialmente dão sentido às frases ou intensificam ou diminuem sentimentos. Dado que estes modelos utilizam a análise direta das palavras individuais e estrutura das frases (conjuntos de palavras e pontuação), sem necessidade de uma representação numérica dos comentários, existe uma maior liberdade sobre o tratamento dos dados. O resultado do pré-processamento transforma os comentários em listas de *tokens* de palavras e também de sinais de pontuação, caracteres especiais e números dado que não são removidos aqui.

O modelo não utiliza apenas a correspondência direta entre as palavras e a polaridade correspondente obtida pelo dicionário de sentimentos desenvolvido (capítulo 4) para obter a polaridade de um comentário. A deteção de expressões comummente utilizadas de acordo com a natureza dos comentários, a deteção do contexto de cada comentário, a identificação e o tratamento de modificadores de sentimentos como negações, os intensificadores de polaridade, permitem inferir a polaridade não conhecida de uma palavra ou frase, pela forma de escrita utilizada com recurso a conjunções adversativas e ainda com a deteção de aspetos positivos e negativos expostos nos comentários, bem como permitem que a polaridade obtida pelo modelo seja calculada de acordo com a forma de escrita utilizada e não pela simples associação de uma polaridade pré-definida a cada palavra obtendo, potencialmente, resultados mais exatos.

Para além disto, o resultado do modelo para cada comentário não se limita à identificação da polaridade binária do mesmo mas ainda à intensidade e à magnitude dessa polaridade e, naturalmente, o contexto em que o comentário se insere bem como os aspetos positivos e negativos nele identificados. Este nível de informação dada como resultado seria muito difícil obter num modelo de aprendizagem automática supervisionado, dada a necessidade de um grande conjunto de todo o tipo de dados anotados e menor flexibilidade de implementação. Todas estas fases constituintes do modelo são demonstradas nas próximas secções.

6.1.1 Deteção de Expressões Comuns

Qualquer linguagem possui um conjunto de expressões globais comummente utilizadas. O mesmo acontece em contextos específicos. Estas expressões podem ser constituídas por uma ou mais palavras. No primeiro caso é considerada como uma palavra dependente do contexto, o segundo caso é mais difícil de identificar sem conhecimento prévio das expressões.

Numa solução de análise de sentimentos baseada em dicionários estas expressões podem representar um grande impacto na definição da polaridade de uma ou mais palavras. Como é comum, numa solução baseada em dicionários, o modelo desenvolvido processa cada palavra do comentário de forma sequencial e atribui uma pontuação a cada uma, de acordo com a polaridade definida no dicionário (informação mais detalhada sobre este processo é apresentada em 6.1.3). Isto leva a que uma expressão comum não detetada ou tratada influencie significativamente a polaridade como é demonstrado no exemplo abaixo.

Exemplo 1: Considere-se que a uma palavra positiva é atribuída a pontuação *+1*, negativa *-1* e neutra *0* e ainda o comentário cuja polaridade é claramente negativa:

"O dispositivo deixa muito a desejar. Às vezes o barato sai caro."

Após o pré-processamento o comentário é transformado numa lista de *tokens* e são removidas as *stop words*, entre outras transformações. Considere-se a seguinte pontuação atribuída ao comentário pré-processado, sem reconhecimento de expressões comuns:

dispositivo	deixa	muito	desejar	.	vezes	barato	sai	caro	.
0	0	0	0	-	0	+1	0	-1	-

A polaridade do comentário é calculada pela soma das pontuações de cada *token*. Uma pontuação final negativa revela um considerado negativo e uma pontuação positiva (incluindo 0) revela um comentário positivo. A polaridade do comentário seria então igual a 0 – positivo – polaridade não correta.

Considere-se agora as duas expressões comuns, frequentemente utilizadas no contexto de análise e presentes no comentário:

- *"deixa muito a desejar"*
- *"barato sai caro"*

O reconhecimento destas expressões transforma o comentário e a pontuação obtida em cada *token* da forma abaixo:

dispositivo	deixa	muito	desejar	.	vezes	barato	sai	caro	.
0	-1			-	0	-1			-

A polaridade obtida com o reconhecimento destas expressões passa a ser igual a -2 ((-1) + (-1)) – negativo – que é de facto a polaridade correta do comentário.

As expressões comuns devem, então, ser consideradas e tratadas como uma única palavra ou *token*, sendo a respetiva polaridade é facilmente obtida pelo dicionário de sentimentos como sendo apenas uma única palavra. A deteção destas expressões é, então, efetuada após o pré-processamento dos comentários e transforma as listas de *tokens* de cada comentário em novas listas sempre que uma expressão comum seja identificada pela unificação dos *tokens* da mesma num único *token*.

As expressões comuns que o modelo é capaz de identificar foram obtidas pela análise extensiva do conjunto de dados, pela forma como as opiniões são comummente expressas. Foi, desta forma, desenvolvido um conjunto de 123 expressões comuns que são procuradas em cada comentário. A tabela 6.1 mostra algumas destas expressões.

Expressões Comuns (exemplos)
Deixa muito desejar
Barato sai caro
Sistema operativo
Não é grande coisa
Cinco estrelas
Gasta pouco
Bom/Boa dia/tarde/noite
Baixo ruído
Não esperava tanto por tão pouco
Só é pena

Tabela 6.1: Exemplos de expressões comuns.

6.1.2 Deteção do Contexto

A deteção do contexto de um comentário serve dois propósitos principais. O mais relevante está relacionado com a diferença de polaridade que uma mesma palavra pode ter em diferentes contextos, fazendo com que seja necessário conhecer à priori o contexto para a escolha da polaridade a aplicar. Uma vez determinado o contexto de um comentário esta informação serve também como resultado adicional à polaridade, acrescentando mais informação ao resultado final.

O contexto de um comentário é definido de acordo com o tipo de produto nele descrito. Os contextos gerais dos produtos do conjunto de dados já foram apresentados na tabela 3.1 do capítulo 3. Considere-se agora a correspondência abaixo para melhor compreensão dos possíveis contextos dos vários produtos.

Produto	Contexto
Máquina lavar roupa/louça	Eletrodomésticos
Aquecedores	Eletrodomésticos
Frigoríficos	Eletrodomésticos
Colunas	Som
Smartphones	Comunicações
Câmaras fotográficas	Fotografia
Computadores	Informática
Monitores	Imagem
TV's	Imagem

Para cada comentário deve então ser obtido o contexto de acordo com o tipo de produto de análise. Inicialmente foi considerado e desenvolvido um modelo de clustering *K-Means* para obtenção desta informação, no qual os comentários de um mesmo contexto pertenceriam ao mesmo cluster. A utilização deste tipo de clustering não se mostrou eficaz na correta separação dos contextos. Foi, então, necessário a utilização de um modelo específico e adequado para a modelação de tópicos/contextos em conteúdos textuais – o *Latent Dirichlet Allocation LDA*.

O *LDA* é um modelo de aprendizagem automática não-supervisionado que identifica automaticamente tópicos presentes em conteúdos textuais pelo reconhecimentos de padrões de co-ocorrência de palavras. Os tópicos podem, então, ser definidos como padrões de co-ocorrência de termos ou palavras num corpus. Por exemplo, o resultado de um modelo *LDA* para o tópico ou contexto *"saúde"* deveria ser *"doutor"*, *"paciente"*, *"hospital"*. A aplicação de *LDA* para identificação do contexto de cada comentário, segue então o princípio de dado o conjunto de tópicos ou contextos conhecidos, atrás apresentados, o *LDA* constrói um conjunto de termos

ou palavras características desses tópicos ou contextos, permitindo associar cada comentário a um tópico ou contexto, tornando também possível a aplicação a futuros novos comentários para determinar o respetivo contexto.

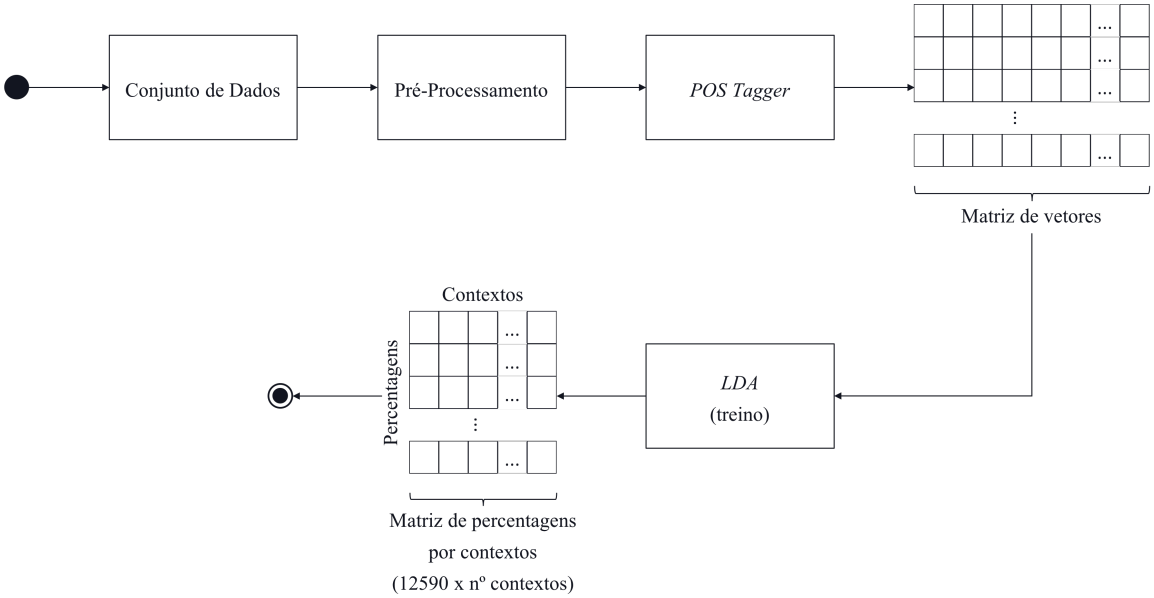


Figura 6.2: A arquitetura do modelo *LDA*.

Como em qualquer outro modelo, os dados foram pré-processados com as mesmas transformações que foram aplicadas no pré-processamento do modelo baseado em dicionários (do qual *LDA* é parte integrante) e também com as expressões comuns já identificadas (ver secção anterior). Após o pré-processamento foi aplicado o *POS Tagger* desenvolvido, para marcar as palavras com as respetivas partes do discurso. Foram mantidas todas as palavras consideradas nomes/substantivos e verbos, o que reduziu o vocabulário. Estas partes do discurso foram utilizadas, uma vez que eram as mais representativas de contextos. Os comentários foram depois representados numericamente numa matriz de vetores, de forma idêntica à realizada nos modelos anteriores. Porém, o vocabulário foi reduzido ainda mais pela definição de um limite superior de frequência de palavras entre os comentários – todas as palavras que ocorram em mais de 80% de todos os comentários são descartadas do vocabulário. A definição deste limite teve como base o facto de que uma palavra que ocorra em todos ou numa grande maioria de todos os comentários, não consegue distinguir o contexto, introduzindo apenas ruído no processo de aprendizagem.

A tabela 3.1 do capítulo 3 mostra os contextos ótimos dos comentários (8). O treino de *LDA* para identificação desses 8 contextos obteve resultados muito pouco satisfatórios, devido à dificuldade de distinção entre os contextos de certos produtos. Os produtos com contextos *comunicações*, *foto*, *imagem*, *som* e *informática* são especialmente difíceis de distinguir dada

a proximidade elevada dos componentes constituintes dos respetivos produtos e dos termos utilizados para os caracterizar (ex.: *foto*, tanto pode ser utilizado para smartphones como para câmaras fotográficas, e *qualidade de som*, tanto pode ser utilizado para smartphones, colunas, computadores). Este facto levou a uma redução da quantidade de possíveis contextos para apenas dois:

- Eletrodomésticos
- Eletrónicos

O modelo foi, então, treinado para identificar os dois possíveis contextos acima. O resultado é uma matriz 12590 x 2 (nº de comentários x nº de contextos), na qual a cada contexto está associada uma percentagem de o comentário pertencer a esse contexto. A tabela 6.2 mostra os valores obtidos para os primeiros 5 comentários do conjunto de dados.

Contexto 0	Contexto 1	Contexto considerado
0.75	0.25	0
0.87	0.13	0
0.83	0.17	0
0.49	0.51	1
0.54	0.46	0

Tabela 6.2: Contextos (percentagens) dos primeiros comentários do conjunto de dados.

Embora não seja possível avaliar com exatidão os resultados do modelo, dado que não existe informação no conjunto de dados do contexto de cada comentário, é possível avaliar manualmente o comportamento e as escolhas do modelo. A tabela 6.3 mostra as primeiras palavras mais relevantes utilizadas pelo modelo em cada contexto e a tabela 6.4 a quantidade de comentários atribuídos a cada contexto.

Contexto 0	máquina	ruído	capacidade	aspirador	limpeza	frigorífico	barulho	lavagem
Contexto 1	som	imagem	design	funcionalidades	câmara	monitor	tablet	software

Tabela 6.3: Primeiras palavras mais relevantes por contexto utilizadas por *LDA*.

Contexto	#Comentários
0 (eletrodomésticos)	7424
1 (eletrónicos)	5166

Tabela 6.4: Quantidade de comentários por contexto.

A análise manual dos resultados permitiu concluir que o contexto 0 corresponde ao contexto dos eletrodomésticos e o contexto 1 ao dos eletrónicos. A quantidade de comentários por contexto é aceitável, embora não seja possível saber, com certeza, se a associação de um comentário a um contexto é a correta. A análise geral da atribuição do contexto a cada comentário mostrou que o modelo não é perfeito, mas os resultados obtidos são aceitáveis, tendo um efeito positivo quando inserido no modelo baseado em dicionários na atribuição de polaridade a palavras específicas de contexto. A atribuição da polaridade a este tipo de palavras e às restantes será apresentado na próxima secção.

6.1.3 Polaridade das Palavras

O pré-processamento dos dados, a deteção de expressões comuns e a deteção do contexto transformam os comentários num formato e informação apropriados para aplicação do dicionário de sentimentos, como forma de obtenção da polaridade das palavras constituintes dos comentários. Aqui é, então, aplicada/obtida a polaridade das palavras dependentes do contexto e das restantes palavras, nesta ordem. O processo está representado na figura 6.3.

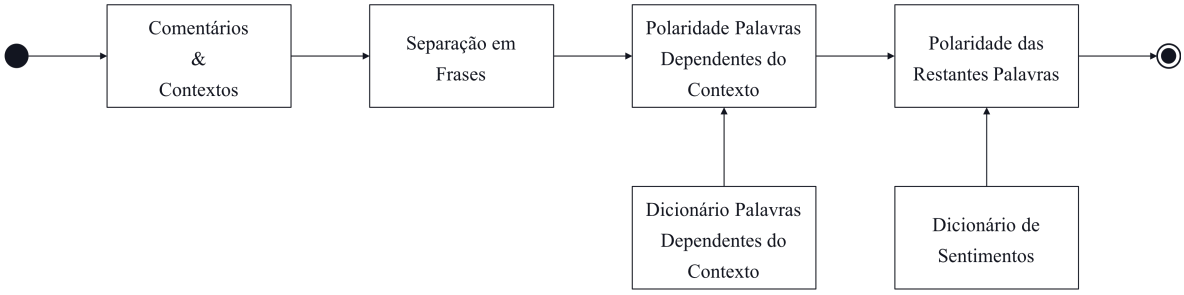


Figura 6.3: Processo de obtenção da polaridade das palavras.

Assim, os comentários começam por ser separados em frases. Esta separação para além de permitir conseguir o nível da frase, permite que um comentário seja tratado em várias partes individuais, representativas de aspetos distintos presentes nos comentários. Muitas vezes a separação de um documento de análise nas suas frases constituintes não é um processo trivial. A utilização de pontos finais, como marcas de separação, é a forma mais simples de separação em frases, mas nem sempre é propicio porque pode facilmente quebrar o sentido das frases individuais – distinção entre pontos finais e uma sigla e não garantia que a frase seguinte a um ponto final não faz qualquer referência a aspetos da frase anterior. A separação dos comentários em frases individuais foi efetuada em:

- *Pontos finais*
- *Vírgulas*

A natureza dos comentários permite que sejam utilizadas estas pontuações para separação em frases individuais, dado que são escritos de forma simplista. A presença de siglas não levanta qualquer problema nesta separação dado que no pré-processamento as mesmas foram identificadas e consideradas como uma única palavra. Comummente, um aspeto é avaliado até a um ponto final, sendo a informação seguinte a esse ponto referente a um novo aspeto ou ao produto em geral. De igual forma, aspetos distintos tendem, também, a ser analisados entre vírgulas. Estes factos permitem justificar a viabilidade destas pontuações para separação dos comentários em frases individuais. Note-se que esta separação em frases é especialmente utilizada como forma de separação de diferentes aspetos, servindo como um passo intermédio para a obtenção dos mesmos e das respetivas polaridades (nível do aspeto). Considere-se o seguinte comentário do conjunto de dados com contexto *eletrodomésticos* e a sua representação em lista de tokens:

"Grande variedade de programas, fácil utilização, pouco ruído, duração dos ciclos um pouco longa. No geral é uma boa compra"

['grande', 'variedade', 'programas', ',', 'fácil', 'utilização', ',', 'pouco', 'ruído', ',', 'duração', 'ciclos', 'um pouco', 'longa', '.', 'geral', 'é', 'boa', 'compra']

A separação em frases transforma o comentário numa lista de listas de tokens:

[['grande', 'variedade', 'programas', ','], ['fácil', 'utilização', ','], ['pouco', 'ruído', ','], ['duração', 'ciclos', 'um pouco', 'longa', '.'], ['geral', 'é', 'boa', 'compra']]

A polaridade de uma palavra é definida numericamente com a mesma representação utilizada no dicionário de sentimentos:

- -1 – negativa
- 1 – positiva

Para cada palavra de cada frase é utilizada a correspondência entre a polaridade definida nas palavras dependentes do contexto e no dicionário de sentimentos. As palavras reconhecidas são alteradas para a respetiva polaridade (palavras neutras não são alteradas e palavras que sejam modificadores de sentimentos – sentiment shifters – também não, de modo a não invalidar o seu efeito na polaridade da palavra que modifica (ver secção seguinte)). Assim, primeiro obtem-se a polaridade das palavras dependentes do contexto para que não existam conflitos com a polaridade definida no dicionário de sentimentos numa mesma palavra, que também seja dependente do contexto. As palavras dependentes do contexto foram obtidas de forma manual, por análise dos comentários sem aplicação de deteção de contextos. A análise efetuada permitiu desenvolver um pequeno dicionário de palavras de sentimento dependentes do

contexto onde para cada palavra está associada a polaridade em cada um dos contextos. Tal como no dicionário de sentimentos, o dicionário de palavras dependentes do contexto pode ser atualizado ao longo do tempo. O dicionário é constituído por 22 palavras, algumas das quais são apresentadas na tabela 6.5.

Palavra	Eletrodomésticos	Eletrónicos
Aquece	1	-1
Silenciosa/o (as/os)	1	-1
Longa/o (as/os)	-1	1
Negro	-1	0
Pequena/o (as/os)	-1	0
Seca	1	0
Vibra	-1	0
Profundidade	1	0

Tabela 6.5: Dicionário de palavras dependentes do contexto (extrato).

O comentário anteriormente apresentado contém uma palavra considerada dependente do contexto: *longa*. A representação final do comentário depois de atribuída a polaridade às palavras fica:

[['grande', **1.0**, 'programas', ','], [**1.0**, 'utilização', ','], ['pouco', **-1.0**, ','], ['duração', 'ciclos', 'um pouco', **-1.0**, '.'], ['geral', 'é', **1.0**, 'compra']]

A polaridade de um comentário é obtida pela soma das polaridades das suas frases. Desta forma é possível lidar com frases com mais do que uma opinião, especialmente se estas possuírem opiniões com polaridades contrárias, tal como acontece no exemplo 1, da secção 2.6.2. Este método permite também classificar subjetivamente cada frase, na qual uma frase totalmente constituída por apenas palavras não é subjetiva, ou seja, não exprime qualquer sentimento ou opinião, sendo suficiente a existência de apenas um valor numérico na representação da frase para que seja subjetiva. O comentário utilizado como exemplo é positivo e a soma das polaridades das frases é também positiva. Porém, isto nem sempre acontece sem utilização de outros recursos como a identificação e o tratamento de modificadores de sentimentos. De facto, atente-se na polaridade da terceira frase – *pouco ruído*. Esta é uma frase claramente positiva, mas a polaridade determinada nesta fase é negativa. A próxima secção mostra o motivo para tal e como essa polaridade é transformada em positiva.

6.1.4 Modificadores de Sentimentos (*sentiment shifters*)

Embora uma palavra tenha uma (ou várias) polaridade associada, dependendo do contexto onde se insere, essa polaridade não é final ou imutável. A polaridade de uma palavra de sentimento pode ser facilmente modificada pela presença de modificadores de sentimentos. Estes modificadores são comummente palavras ou expressões que quando associadas a uma palavra de sentimento alteram a polaridade dessa palavra. Os modificadores de sentimentos mais frequentes são palavras de negação, mais concretamente, advérbios de negação como *não*, *tampouco*, *nem*, *nunca*, que invertem a polaridade de uma palavra ou de uma frase quando presentes. Outras palavras e expressões, e até formas verbais, podem ser modificadores de sentimentos. Como tal, devem também ser identificados nos documentos de análise. A figura 6.4 mostra o esquema simplificado da identificação de modificadores de sentimentos e inversão da polaridade por eles provocada.

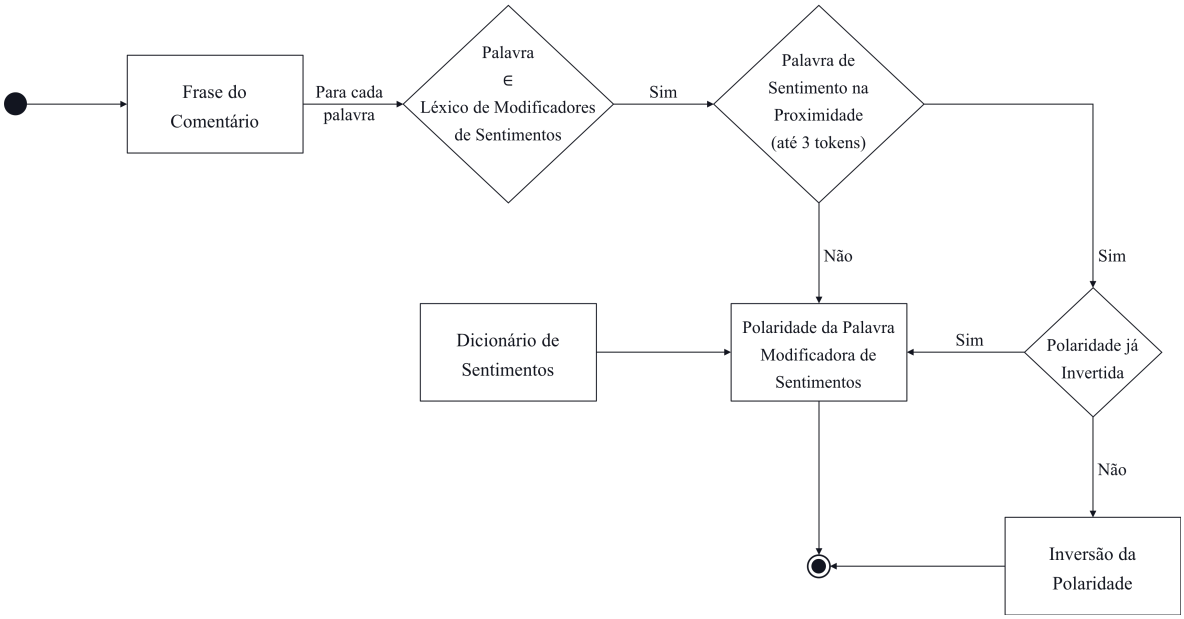


Figura 6.4: Processo de identificação de modificadores de sentimentos e inversão da polaridade.

No início desta fase do modelo baseado em dicionários, os comentários estão separados em frases individuais e representados em listas de listas de tokens, em que alguns dos tokens são valores numéricos representativos da polaridade já obtida das palavras que substituem, como foi visto no final da secção anterior. A identificação de sentiment shifters é efetuada palavra a palavra, para cada frase do comentário. É, então, verificado para cada palavra se a mesma pertence ao léxico de modificadores de sentimentos desenvolvido.

O léxico de modificadores de sentimentos desenvolvido é constituído por um conjunto de palavras, algumas das quais são de negação comuns e advérbios de negação, outras foram

obtidas pela análise manual dos comentários, o que permitiu também identificar forma verbais comummente utilizadas para este efeito. O léxico é constituído por 28 palavras como mostra a tabela 6.6.

Podia	Nunca	Apenas	Deixa
Pouco	Jamais	Sem	Deixou
Pouca	Poderia	Espera	Deve ser
Nada	Seria	Esperava	Devia ser
Não	Deveria	Impedindo	Fosse
Tampouco	Deveriam	Impede	Resolve
Nem	Devia	Menos	Abaixo

Tabela 6.6: Modificadores de sentimentos (*sentiment shifters*) – léxico.

Uma vez identificada uma palavra do comentário como um sentiment shifter é verificado se na sua proximidade existe uma polaridade, ou seja, é utilizada a proximidade de uma polaridade a uma palavra modificadora de sentimentos como método de inversão dessa polaridade. Este processo assenta no facto de quando uma palavra modificadora de sentimentos ocorre na proximidade de uma palavra de sentimento, então a polaridade desta última é afetada pela primeira. O método utilizado em [11] utiliza o mesmo princípio. A análise dos comentários permitiu definir a proximidade igual a 3 tokens subsequentes ao sentiment shifter. Neste tipo de situações, podem acontecer dois casos distintos:

1. **Polaridade existe na proximidade.**
   Caso exista uma polaridade na proximidade do sentiment shifter então essa polaridade é candidata a ser afetada, ou seja, invertida. Uma polaridade é invertida apenas se não tiver sido invertida anteriormente por um outro sentiment shifter, garantindo assim que uma polaridade é invertida apenas uma vez. Caso a polaridade não tenha sido anteriormente alterada, a mesma é invertida sendo efetuado o processo para a próxima palavra da frase em análise. Naturalmente que, a inversão de uma polaridade positiva (1.0) resulta numa polaridade negativa (-1.0) e vice-versa. Caso a polaridade já tenha sido alterada por um modificador de sentimentos então não é invertida e é aplicado o ponto 2.

2. **Polaridade não existe na proximidade/Polaridade existe na proximidade mas já foi invertida.**
   Caso não exista uma polaridade na proximidade do sentiment shifter ou exista mas essa polaridade já tenha sido previamente alterada por um outro sentiment shifter, então a palavra deixa de ser considerada um modificador de sentimentos e passa a ser considerada uma palavra de sentimento, podendo agora ser utilizado o dicionário de sentimentos para

tentar determinar a sua polaridade.

Apesar de simples, este processo de deteção de modificadores de sentimentos e de inversão das polaridades afetadas mostrou-se eficaz sobre o contexto de análise.

Considere-se o comentário exemplo utilizado na secção anterior e a sua representação final nessa secção.

[['grande', **1.0**, 'programas', ','], [**1.0**, 'utilização', ','], ['pouco', **-1.0**, ','], ['duração', 'ciclos', 'um pouco', **-1.0**, '.'], ['geral', 'é', **1.0**, 'compra']]

Como referido anteriormente, a terceira frase possui uma polaridade contrária à que é expressa. Com a identificação de modificadores de sentimentos é possível obter a polaridade correta da frase – a palavra *pouco*, presente na frase, é considerada um sentiment shifter (tabela 6.6). Como existe uma polaridade ainda não invertida na sua proximidade, essa polaridade é, então, invertida obtendo-se uma polaridade positiva na frase que é, de facto, a polaridade correta, como se mostra abaixo.

[['grande', **1.0**, 'programas', ','], [**1.0**, 'utilização', ','], ['pouco', **1.0**, ','], ['duração', 'ciclos', 'um pouco', **-1.0**, '.'], ['geral', 'é', **1.0**, 'compra']]

A título adicional de exemplo considere-se o seguinte comentário do conjunto de dados com polaridade negativa e com a representação do comentário antes e depois da aplicação de modificadores de sentimentos:

"O tapete poderia ser um pouco mais alto seria perfeito e ter mais cores à escolha."

[['tapete', 'poderia', 'ser', 'um pouco', 'mais', 'alto', 'seria', **1.0**, 'ter', 'mais', 'cores', 'escolha', '.']]

[['tapete', **-1.0**, 'ser', 'um pouco', 'mais', 'alto', 'seria', **-1.0**, 'ter', 'mais', 'cores', 'escolha', '.']]

Sem a aplicação de sentiment shifters a polaridade do comentário seria positiva (polaridade errada) mas esta transformação consegue obter a polaridade correta. Duas palavras modificadoras de sentimentos estão presentes no comentário: *poderia* e *seria*. A última altera a polaridade que se encontra na proximidade (de modo igual ao que sucedeu no comentário anterior). A primeira, no entanto, não possui qualquer polaridade na sua proximidade e como tal passa a ser tratada como uma palavra de sentimento, sendo assim utilizado o dicionário de sentimentos para obter a sua polaridade que, devido à forma verbal, implica a necessidade de algo que não acontece no produto e como tal é considerada negativa.

6.1.5 Intensificadores de Sentimentos

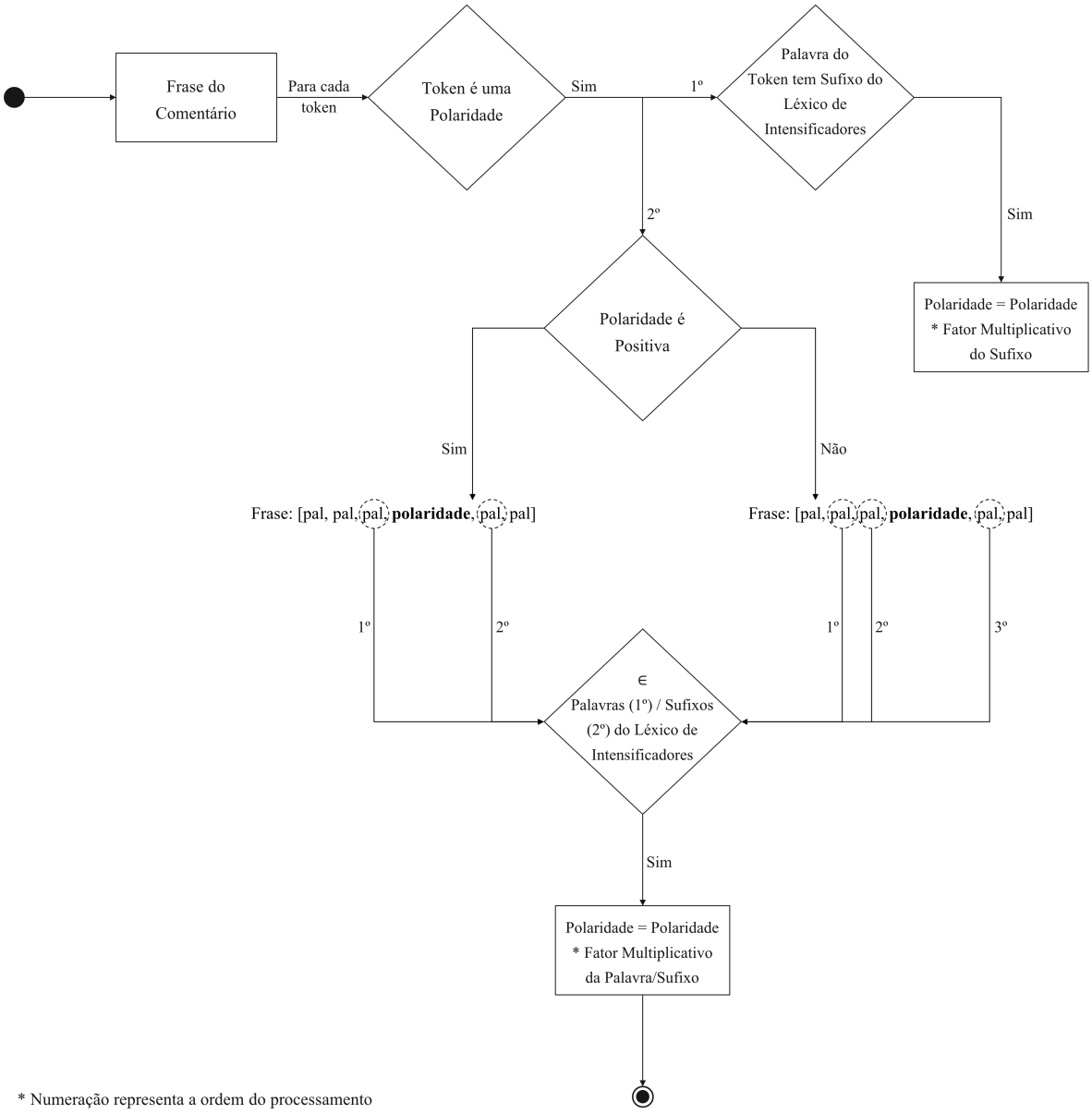


Figura 6.5: Processo de intensificação das polaridades/sentimentos.

Até este ponto, a polaridade das palavras tem sido representada numericamente por 1.0 (positiva) e -1.0 (negativa). Esta representação apesar de ser eficaz para distinguir as duas polaridades, não consegue de qualquer forma distinguir entre a mesma polaridade. Uma mesma polaridade é distinguida entre si pela intensidade que é exprimida pelas palavras, por exemplo, *gostei* e *gostei muito* são duas frases positivas mas a última expressa um grau de positividade maior – o mesmo acontece entre as palavras *bom* e *ótimo* onde a intensidade da última é mais

elevada. A intensidade dos sentimentos expressos num documento é dada pelas palavras de intensidade. A identificação deste tipo de palavras e associação a uma polaridade existente permite que uma análise de sentimentos em textos seja mais realista e com resultados que permitem comparar não só entre as polaridades binárias mas também entre a mesma polaridade. Neste sentido, foi desenvolvido um sistema de identificação de intensificadores sobre polaridades conhecidas que faz uso de um léxico de intensificadores de sentimentos. A figura 6.5 mostra o esquema geral deste sistema. A sua descrição e a dos seus constituintes está apresentada de seguida.

O sistema faz uso de um léxico de intensificadores de sentimentos desenvolvido pela análise dos comentários e de um conjunto de palavras e formas verbais características de intensidade. Advérbios de quantidade, grau e intensidade são as palavras que mais frequentemente exprimem intensidade. Para além de palavras deste tipo, outras palavras e formas verbais foram utilizadas para construção do léxico como certos sufixos que quando aplicados a uma palavra são tratados como um intensificadores. Um dos sufixos mais comuns é *"mente"*, utilizado em palavras como *"totalmente"* e *"extremamente"*.

Palavra	Fator_Multiplicativo	Sufixo	Fator_Multiplicativo
bastante	1.8	mente	2
demais	1.6	íssima	2
demasiada	1.6	íssimo	2
demasiado	1.6	íssimas	2
demasiadas	1.6	íssimos	2
demasiados	1.6	érrima	2
deveras	1.6	érrimo	2
muita	1.8	érrimas	2
muito	1.8	érrimos	2
muitas	1.8	ima	2
muitos	1.8	imo	2
tanto	1.2	imas	2
tão	1.4	imos	2

Tabela 6.7: Intensificadores de sentimentos – léxico.

O léxico de intensificadores é, portanto, constituído por um conjunto de palavras e de sufixos. Dado que se pretende alterar a positividade ou negatividade de uma polaridade existente, foi utilizado um fator multiplicativo que foi aplicado sobre a polaridade. A cada palavra e sufixo do léxico está, então, associado o respetivo fator multiplicativo que representa o grau de intensidade que a palavra exerce sobre a polaridade. O léxico é constituído por 26

entradas e o fator multiplicativo de cada uma foi escolhido dentro do intervalo de valores [1.2, 2] em escala de 0.2. A tabela 6.7 mostra o léxico desenvolvido. O fator de multiplicidade escolhido para cada palavra ou sufixo permite distinguir, por exemplo, *"muito satisfeito"* de *"totalmente satisfeito"*, onde a última possui um fator de multiplicidade (intensidade) maior. De facto, todos os sufixos possuem o maior nível de multiplicidade possível. Estes fatores de multiplicidade levam a que uma polaridade positiva passe agora a estar dentro do intervalo [1.0, 2.0] em escala/intervalos de 0.2 e o mesmo acontece para polaridades negativas dentro do mesmo intervalo de valores negativos.

A identificação de intensificadores e a aplicação do fator de multiplicidade segue o esquema da figura 6.5 e, tal como aconteceu nos modificadores de sentimentos/sentiment shifters da secção anterior, o processo baseia-se na proximidade destes intensificadores em relação à polaridade. O processo é efetuado para cada token de cada frase de um comentário. Se o token for uma polaridade, anteriormente obtida, então o processo é realizado pela seguinte ordem:

1. Se a palavra que deu origem à polaridade tiver um dos sufixos do léxico de intensificadores de sentimentos então é aplicado o respetivo fator multiplicativo sobre a polaridade. Esta procura sobre a palavra representada pela polaridade permite distinguir entre casos em que esta própria palavra é um intensificador. Considere-se, por exemplo, a distinção entre *"bom"* e *"ótimo"* nos pseudo comentários e a sua representação em listas de tokens com polaridade identificada:

C1: *O produto é bom* => [['o', 'produto', 'é', 1.0 (bom)]]

C2: *O produto é ótimo* => [['o', 'produto', 'é', 1.0 (ótimo)]]

Polaridade C1 = C2 = 1.0

A identificação de intensificadores transforma a representação dos comentários em:

C1: [['o', 'produto', 'é', 1.0 (bom)]]

C2: [['o', 'produto', 'é', 2.0 (ótimo)]]

Polaridade C1 = 1.0
Polaridade C2 = 2.0

C1 mantém a representação e a polaridade, mas C2 possui agora uma maior polaridade devido à intensidade superior da palavra *"ótimo"*, em comparação com *"bom"*, dada pela presença do sufixo considerado intensificador com multiplicidade 2x.

2. Se a polaridade é:

- Positiva
  A análise dos comentários permitiu observar que para polaridades positivas as palavras intensificadoras encontram-se em grande frequência imediatamente antes ou depois da polaridade. Assim, se o token ou a palavra imediatamente anterior ou imediatamente a seguir (nesta ordem) à polaridade pertencer às palavras ou sufixos do léxico de intensificadores é aplicado o respetivo fator multiplicativo sobre a polaridade.

- Negativa
  Para polaridades negativas os intensificadores podem ocorrer em posições mais distantes anteriores às polaridades. Neste sentido, se o segundo ou o primeiro token ou a palavra imediatamente anterior ou o token ou a palavra imediatamente a seguir à polaridade (nesta ordem) pertencer às palavras ou sufixos do léxico de intensificadores é aplicado o respetivo fator multiplicativo sobre a polaridade.

Como forma de exemplo adicional de aplicação de intensificadores, considere-se os dois comentários do conjunto de dados e as suas representações imediatamente antes e depois da aplicação de intensificadores:

C1: *"Comprei esta semana, são extremamente desconfortáveis e inflexíveis."*

[['comprei', 'esta', 'semana', ','], ['são', 'extremamente', **-1.0**, **-1.0**, '.']]

[['comprei', 'esta', 'semana', ','], ['são', 'extrema<u>mente</u>', **-2.0**, **-2.0**, '.']]

C2: *"Muito satisfeito com a compra deste tablet. Aconselho vivamente."*

[['muito', **1.0**, 'compra', 'deste', 'tablet', '.'], [**1.0**, 'vivamente', '.']]

[['<u>muito</u>', **1.8**, 'compra', 'deste', 'tablet', '.'], [**2.0**, 'viva<u>mente</u>', '.']]

No primeiro comentário observa-se a aplicação de um intensificador a duas polaridades distintas, no qual uma delas está a dois tokens de distância, o que é possível por ser uma polaridade negativa. O segundo comentário possui polaridades apenas positivas, ocorrendo as duas possibilidades de aplicação de intensificadores, imediatamente antes e imediatamente depois da polaridade. O resultado final mantém o mesmo tipo de polaridade original para cada comentário, mas agora é possível saber o quanto positivo e negativo é um comentário e comparar comentários com a mesma polaridade. Note-se ainda que, tal como referido anteriormente, a polaridade de uma palavra é agora definida entre 1.0 e 2.0, que é um intervalo de valores relativamente pequeno, que faz com que a intensidade entre dois valores próximos (ex.: 1.0

e 1.4) seja já considerada significativa (isto vai ter um papel fundamental para o calculo da magnitude da polaridade que é apresentado posteriormente).

6.1.6 Tonalidade dos Sentimentos

Um documento de análise ou um comentário pode ter várias polaridades referentes a vários aspetos nele descritos, cuja soma dessas polaridades é uma forma de conseguir a polaridade geral do comentário, como tem sido efetuado até este ponto. Contudo, é frequente que, apesar de várias opiniões sobre vários aspetos sejam apresentadas, uma tonalidade geral em relação à polaridade ou ao sentimento global esteja também presente no comentário. Isto pode ser visível no possível comentário:

"Apesar da bateria durar muito pouco, o equipamento é bom."

Este comentário apresenta um aspeto negativo referente à duração da bateria e um aspeto positivo referente ao equipamento em geral. Apesar da presença de uma polaridade negativa, o tom geral que é pretendido mostrar com o comentário é positivo dada a forma como o mesmo está escrito. Até este ponto, a representação do comentário é dada por:

[['apesar', 'bateria', 'durar', 'muito', **-1.8**, ','], ['equipamento', 'é', **1.0**, '.']]

Dada a representação acima, a polaridade final seria igual a -0.8, logo, negativa que não é a polaridade correta. Uma forma de lidar com o tom presente nos comentários é então necessária (quando presente). Na figura 6.6 podemos ver o esquema geral da solução implementada.

A forma como uma tonalidade é expressa é muito dependente das palavras que são utilizadas. No comentário acima utilizado como exemplo, a primeira frase, que é negativa, é considerada como neutra para o tom geral que é dado pela presença da palavra *"apesar"*. Esta palavra tem a capacidade de anular qualquer sentimento ou opinião expressa na primeira parte do comentário, quando o objetivo é a análise geral do mesmo, de facto, caso essa palavra seja excluída todo o comentário perde o sentido original.

A análise dos comentários permitiu observar que quando ocorrem duas polaridades com sinal distinto (uma positiva e outra negativa) e existe uma conjunção adversativa na proximidade de uma das polaridades, a tonalidade geral do sentimento que se quer exprimir é dada pela parte não afetada pela conjunção adversativa, tal como acontece no comentário exemplo onde *"apesar"* é, de facto, uma conjunção adversativa e a parte que ela afeta é considerada como nula. Foi neste sentido que a solução foi desenvolvida. Foi construído um léxico de palavras e conjunções adversativas que devem ser identificadas de modo a anular a parte por elas afetada, que após a identificação é levantado o problema de qual a parte que essa conjunção afeta. Mais

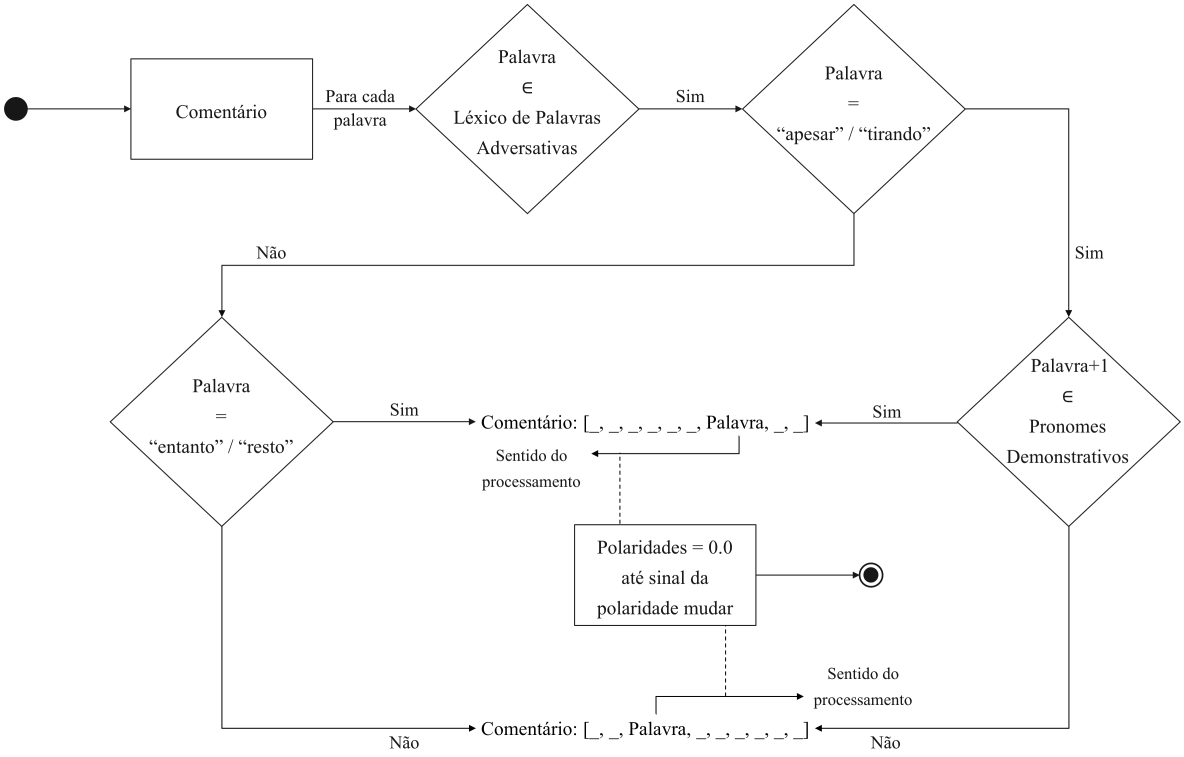


Figura 6.6: Processo de tonalidade dos sentimentos.

especificamente, este problema pode ser dividido em duas partes fundamentais:

- O sentido do processamento
- O ponto de paragem

O sentido do processamento é a decisão da direção a tomar a partir da conjunção adversativa ou seja, se a parte afetada se encontra antes ou depois da conjunção. O método utilizado considera que o processamento é sempre efetuado para a frente da conjunção mas com a ocorrência de exceções que levam à inversão deste sentido. No sentido do processamento as polaridades anteriormente obtidas são, então, anuladas (0.0), até ao ponto de paragem a definir. Considerem-se os seguintes pseudo comentários que, de forma simples, demonstram o processo de decisão do sentido do processamento (conjunções adversativas estão destacadas a negrito e o sentido do processamento/parte ignorada está à frente):

C1: *O produto é bom* ***apesar*** *de ser barulhento.* => Depois

C2: *É barulhento.* ***Apesar*** *disto, gostei do produto.* => Antes

C3: *O produto faz muito ruído.* ***Tirando*** *isto é bom.* => Antes

C4: ***Tirando*** *o ruído, o produto é bom.* => Depois

C5: *O produto é bom. No* ***entanto****, é demasiado caro.* => Antes

C6: ***Embora*** *seja bom, é muito barulhento.* => Depois

C7: *Faz muito ruído mas de* ***resto*** *é bom.* => Antes

A análise cuidada dos comentários permite observar um padrão facilmente utilizado para determinar o sentido. Atente-se apenas nos comentários cujo sentido é para trás da conjunção adversativa (C2, C3, C5, C7) onde em C2 e C3 esta inversão de sentido ocorre pela presença das palavras *"disto"* e *"isto"* que alteram o alvo da conjunção adversativa o qual não acontece em C1 e C4 que são comentários semelhantes. Estas palavras são, de facto, pronomes demonstrativos cujo papel é identificar o alvo do contexto e, quando aplicados às conjunções adversativas *"apesar"* e *"tirando"* o alvo passa a ser antes delas, alterando assim o sentido do processamento. Isto acontece na presença de qualquer pronome demonstrativo. Os comentários C5 e C7 também têm o sentido do processamento para trás da conjunção adversativa, mas sem ocorrência de pronomes demonstrativos que se deve ao facto de a conjunção *"entanto"* e a palavra *"(de) resto"* terem de forma natural o alvo minoritário/de menor relevância antes deles. Dadas estas observações, que se aplicam de uma forma geral em qualquer situação ou comentário e não apenas nos comentários utilizados como exemplo, a definição de sentido do processamento é dada quando o:

- Processamento é efetuado para trás da conjunção adversativa se:
  - Conjunção é igual a *"apesar"* ou *"tirando"* e a palavra imediatamente a seguir pertencer à lista de pronomes demonstrativos considerados.
  - Conjunção é igual a *"entanto"* ou *"(de) resto"*.
- Processamento é efetuado para a frente da conjunção adversativa para todos os outros casos.

A tabela 6.8 mostra o léxico de conjunções adversativas utilizado que inclui também o conjunto de pronomes demonstrativos considerados.

Uma vez definido o sentido do processamento, resta definir o ponto de paragem no qual as polaridades deixam de ser anuladas. Observe-se os comentários C4 e C6, anteriormente apresentados, cujo sentido do processamento é para a frente da conjunção adversativa. Nestes comentários o ponto de paragem é essencial dado que apenas a primeira parte/frase do comentário deve ser ignorada, uma vez que sem um ponto de paragem todo o comentário seria ignorado. A decisão do ponto de paragem não é um processo trivial dado que é necessário

Conjunções Adversativas	Pronomes Demonstrativos	
Apesar	Isso	Essa(s)
Tirando	Aquilo	Disto
Entanto	Este(s)	Deste(s)
Embora	Esta(s)	Desta(s)
Resto	Esse(s)	Isto

Tabela 6.8: Conjunções Adversativas – léxico.

saber até que ponto os aspetos descritos deixam de estar associados ao efeito que a conjunção adversativa tem sobre eles. Considere-se o possível comentário que apesar de ter vários aspetos positivos, a polaridade geral que o autor transmite é negativa:

*"**Apesar** do ecrã ser ótimo, câmara muito boa e com som limpo de boa definição, o produto é muito caro."*

Várias aspetos são afetados pela conjunção adversativa e o ponto de paragem deve ser definido de forma a que apenas as polaridades desses aspetos sejam afetadas. O método de paragem utilizado é baseado nas polaridades, uma conjunção adversativa é utilizada quando existe um contraste entre pelo menos dois aspetos que faz com que as polaridades desses aspetos sejam opostas. Considere-se a representação do mesmo comentário numa lista de tokens, sem separação em frases individuais, e com polaridades obtidas pelos processos até agora implementados:

['**apesar**', 'ecrã', 'ser', **2.0**, ',', 'câmara', 'muito', **1.8**, 'som', **1.0**, **1.0**, 'definição', ',', 'produto', 'é', 'muito', **-1.8**, '.']

As polaridades dos aspetos afetados pela conjunção adversativa são todas positivas passando para negativa quando deixa de ser afetado pela conjunção, validando, assim, o contraste formado pela mesma. A solução passa, então, por verificar o sinal da primeira polaridade que ocorre após a conjunção adversativa no sentido do processamento (previamente determinado) e anula todas as polaridades que ocorram nesse sentido até ocorrer uma polaridade com sinal diferente que já não é anulada. A representação do comentário com a deteção da tonalidade é então:

['**apesar**', 'ecrã', 'ser', **0.0**, ',', 'câmara', 'muito', **0.0**, 'som', **0.0**, **0.0**, 'definição', ',', 'produto', 'é', 'muito', **-1.8**, '.']

Todas as polaridades afetadas foram anuladas, deixando de ter peso no cálculo da polaridade global do comentário, indo de encontro com a tonalidade expressa no comentário que passa a ser negativo como o autor do mesmo pretende demonstrar. Este processo representa, assim,

a última transformação responsável por determinar as polaridades presentes num comentário, sendo a polaridade final dada pelos valores das polaridades obtidas até aqui. Note-se ainda que o léxico de conjunções adversativas não faz uso de todas as conjunções deste tipo consideradas na gramática, mas apenas das mais relevantes de acordo com a análise efetuada. É ainda relevante referir que a passagem das polaridades a 0 não provoca a perda das polaridades originais que são necessárias para o processo de deteção dos aspetos que é apresentado na próxima secção.

6.1.7 Deteção de Aspetos

A flexibilidade de implementação de modelos baseados em dicionários permitiu dirigir a forma de representação dos comentários, de modo a que a identificação dos seus constituintes fosse efetuada da forma o mais simples possível. Isto permite que a identificação dos vários aspetos descritos num comentário seja também possível. O nível do aspeto é o maior nível possível de análise, permitindo obter o conjunto de aspetos que são considerados positivos e negativos, aumentando significativamente o detalhe dos resultados de uma análise.

O método desenvolvido tem por base a proximidade que os aspetos têm em relação a uma polaridade. Uma polaridade descreve sempre um aspeto ou conjunto de aspetos ou ainda o produto em geral, logo, a presença de um aspeto ocorre na sua proximidade. Considere-se os dois comentários abaixo e as respetivas representações em listas de listas (frases) de tokens com polaridades conhecidas e aspetos destacados a negrito:

C1: *"Apesar do ecrã ser ótimo, câmara muito boa e com som limpo de boa definição, o produto é muito caro."*

[['apesar', '**ecrã**', 'ser', 2.0, ','], ['**câmara**', 'muito', 1.8, '**som**', 1.0, 1.0, 'definição', ','], ['produto', 'é', 'muito', -1.8, '.']]

C2: *"Fácil utilização, resistente e excelente autonomia."*

[[1.0, '**utilização**', ','], [1.0, 2.0, '**autonomia**', '.']]

Os comentários acima, embora simples, conseguem mostrar características essenciais para a deteção dos aspetos: todos os aspetos ocorrem, de facto, na proximidade da respetiva polaridade que os define, tanto a polaridade como o aspeto estão presentes na mesma frase e os aspetos podem ocorrer antes ou depois da polaridade.

A análise detalhada à forma como os aspetos ocorrem em relação às polaridades permitiu constatar um comportamento frequente:

- Os aspetos ocorrem com mais frequência antes da polaridade.
- Quando um aspeto está antes da respetiva polaridade pode estar a várias posições de distância da mesma e no máximo até à posição inicial da frase (ex.: ['**câmara**', 'é', 'muito', 'mesmo', 'muito', 'boa (1.8)']).
- Quando um aspeto está depois da respetiva polaridade então ocorre na posição imediatamente a seguir à mesma.

Considere-se agora o seguinte comentário e respetiva representação em listas:

C3: *"Tem excelente imagem, iso, focagem vídeo, etc."*

[['tem', 2.0, '**imagem**', ','], ['**iso**', ','], ['**focagem**', ','], ['**vídeo**', ','], ['etc', '.']]

Ao contrário do que acontece com os comentários C1 e C2, no comentário acima vários aspetos são representados pela mesma polaridade e encontram-se a distâncias cada vez superiores e em frases distintas. Quando uma única polaridade caracteriza vários aspetos, estes ocorrem frequentemente em sequência e separados por vírgulas, que tanto pode ser antes ou depois da polaridade. Este tipo de aspetos são, então, considerados aspetos distantes à respetiva polaridade.

A solução implementada tem por base dois processos principais que direcionam a recolha dos aspetos: sentido da procura dos aspetos em relação à polaridade e procura de aspetos distantes. É utilizado o *POS Tagger* desenvolvido para marcar as palavras com as respetivas partes do discurso para identificação dos aspetos, os aspetos que são explicitamente representados em textos são frequentemente nomes ou substantivos e é, então, utilizada esta parte do discurso como forma de identificação dos mesmos. Isto levou a que fosse desenvolvido um conjunto de palavras cuja parte do discurso é nome (ou substantivo) que apesar de terem esta parte do discurso não podem ser considerados aspetos devido ao alvo que representam como *"máquina"*, *"produto"*, *"telemóvel"*, *"equipamento"*, *"aparelho"*, *"artigo"*, etc., que representam o produto em geral e não um aspeto específico do mesmo. As figuras 6.7 e 6.8 mostram o esquema geral da solução implementada.

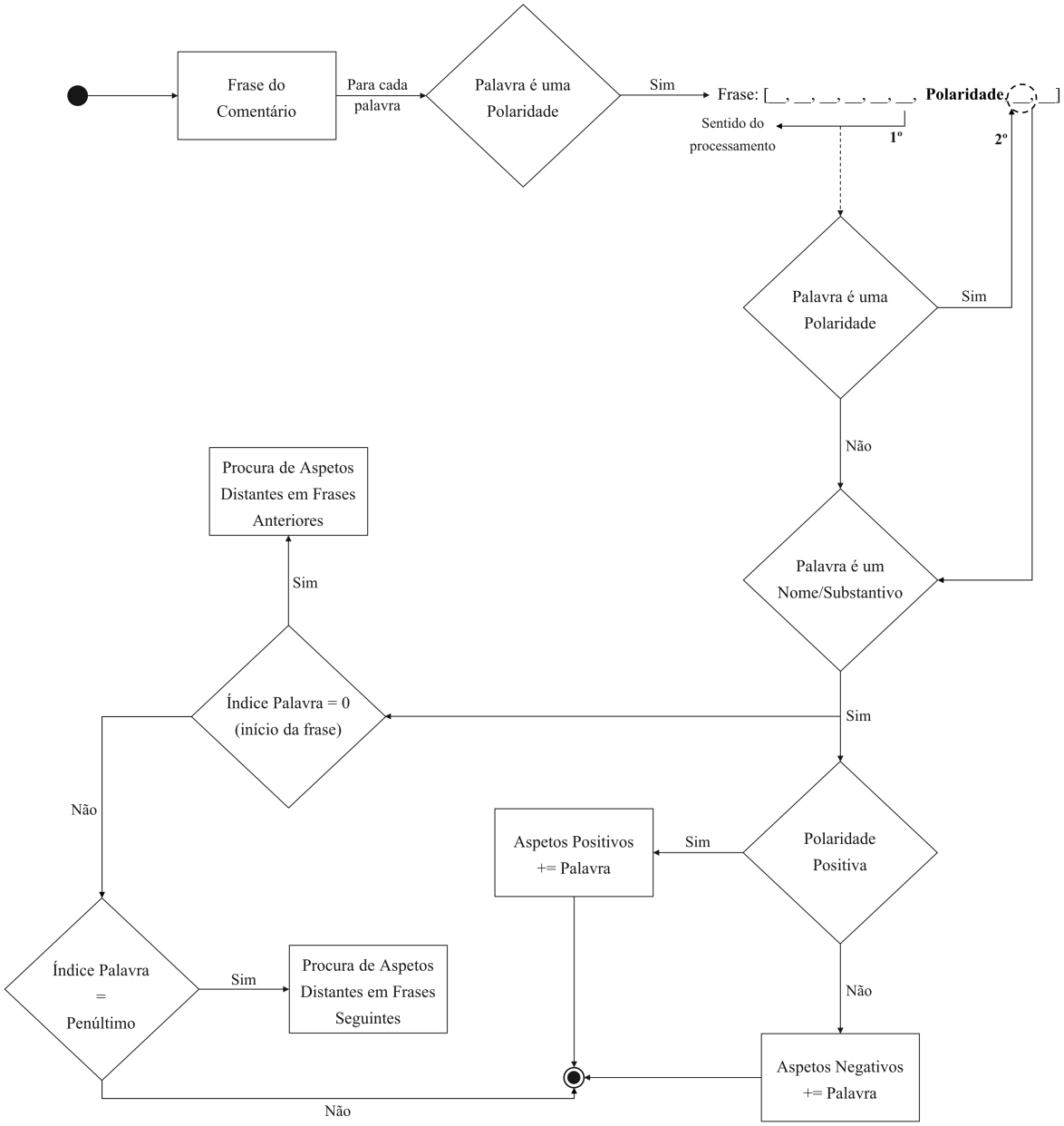


Figura 6.7: Processo de deteção de aspetos.

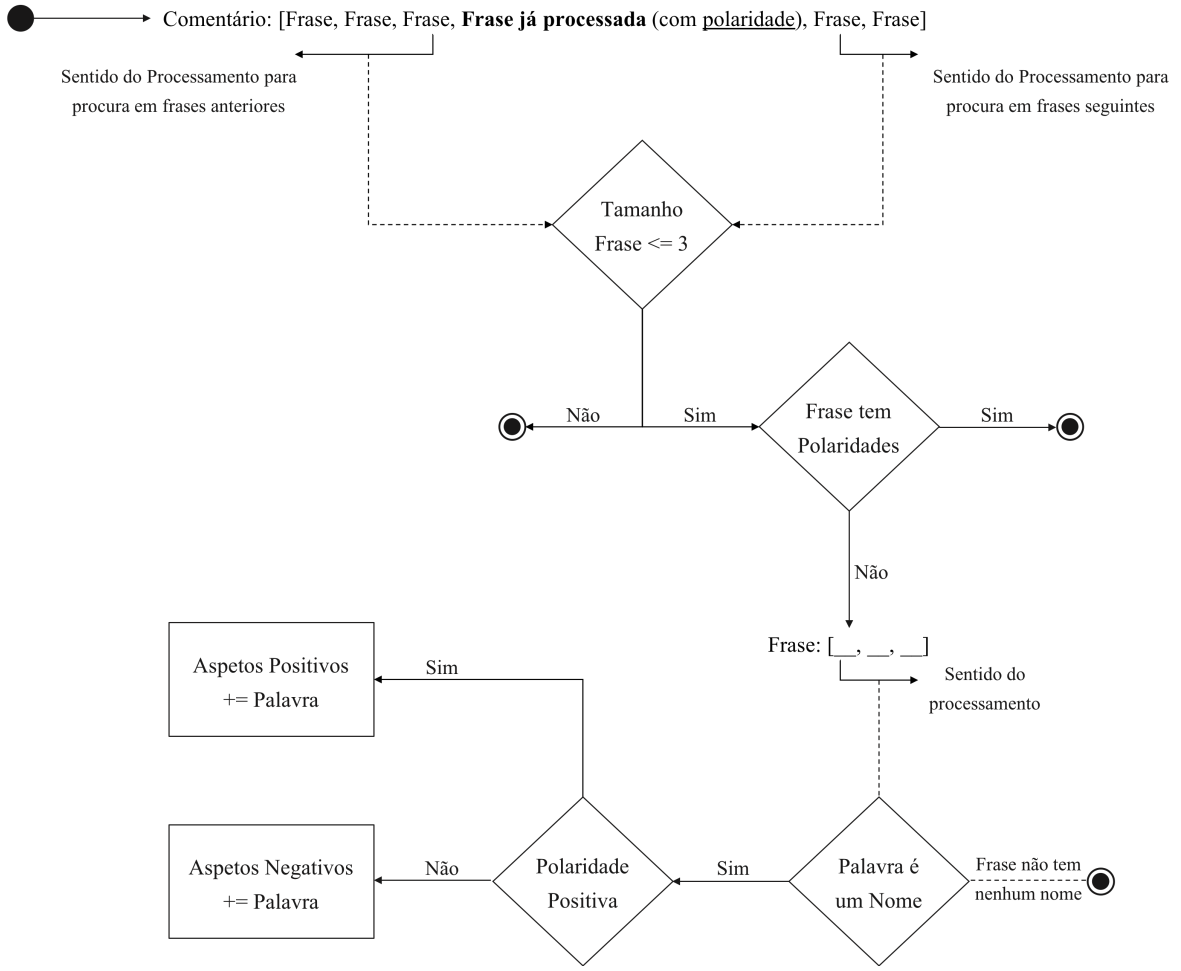


Figura 6.8: Processo de deteção de aspetos distantes em frases anteriores e seguintes.

Os seguintes pontos explicam por palavras o processo de deteção dos aspetos (figura 6.7):

- Para cada palavra de cada frase de um comentário:
  - Se palavra é uma polaridade:
    - Procura para trás da polaridade até ao fim da frase ou até uma nova polaridade:
      - Se palavra é um nome e a polaridade é positiva então a palavra é um aspeto e é adicionada ao conjunto de aspetos positivos.
      - Se palavra é um nome e a polaridade é negativa então a palavra é um aspeto e é adicionada ao conjunto de aspetos negativos.
      - Se palavra é um nome e está na primeira posição da frase então

podem haver aspetos distantes, em frases anteriores, e é efetuada a procura nessas frases.

* Se não foi encontrado um aspeto atrás da polaridade então é procurado à frente da mesma, apenas na posição imediatamente a seguir:

  · Se palavra é um nome e a polaridade é positiva então a palavra é um aspeto e é adicionada ao conjunto de aspetos positivos.

  · Se palavra é um nome e a polaridade é negativa então a palavra é um aspeto e é adicionada ao conjunto de aspetos negativos.

  · Se palavra é um nome e está na penúltima posição da frase (note-se que a última posição tem sempre um sinal de pontuação e não uma palavra) então podem haver aspetos distantes, em frases seguintes, e é efetuada a procura nessas frases.

Os seguintes pontos explicam por palavras o processo de deteção dos aspetos distantes em frases anteriores e seguintes à frase já processada (figura 6.8). É apresentada a explicação para a deteção nas frases anteriores sendo que para as frases seguintes todos os pontos são válidos, mas para a frente da frase já processada:

- Para cada frase do comentário, anterior à frase já processada, desde esta até à frase início:

  – Se o tamanho da frase é menor ou igual a 3 (tamanho = nº tokens/palavras) e a frase não tem qualquer polaridade:

    * Para cada palavra da frase:

      · Se palavra é um nome e a polaridade da frase já processada é positiva (polaridade que define o aspeto) então a palavra é um aspeto e é adicionada ao conjunto de aspetos positivos.

      · Se palavra é um nome e a polaridade da frase já processada é negativa (polaridade que define o aspeto) então a palavra é um aspeto e é adicionada ao conjunto de aspetos negativos.

      · O processo é efetuado para as próximas frases, neste caso anteriores, enquanto os pontos acima se verificarem para cada uma.

A solução desenvolvida consegue, então, identificar o conjunto de aspetos explicitamente presentes nos comentários e tem como resultado uma lista dos aspetos positivos e outra dos

aspetos negativos. O método é simples e muito eficaz, especialmente quando os comentários estão bem escritos, com pontuação e estrutura corretas/adequadas. Para comentários menos cuidados os resultados são, também, bons embora possa ocorrer a eventual não deteção de um aspeto ou incorreta associação a uma polaridade que não o representa e ainda a consideração de um aspeto que na verdade não é, devido à incorreta marcação da parte do discurso como um nome ou substantivo por parte do *POS Tagger* desenvolvido. De facto, todas as transformações implementadas até este ponto assentam umas nas outras que faz com que a ocorrência de um erro numa delas possa comprometer o resultado das seguintes.

6.2 Resultados Finais

Após todo o conjunto de transformações aplicado sobre os comentários, obtiveram-se três tipos de informação principais: contexto de cada comentário, conjunto de aspetos positivos e negativos, que estão explicitamente descritos em cada comentário, e a representação dos comentários em listas de frases com polaridades conhecidas (em formato decimal) que podem ser negativas, neutras ou positivas. Isto permite aumentar significativamente a quantidade de informação final que é apresentada depois do processamento de um comentário, em comparação com a única polaridade binária do modelo supervisionado de aprendizagem automática. O resultado final de uma análise é constituído por:

- Polaridade
- Magnitude
- Contexto
- Aspetos positivos
- Aspetos negativos

Tanto a informação do contexto como dos aspetos foi já obtida ao longo das transformações que no caso do contexto é utilizada para determinar a polaridade das palavras dependentes do contexto. Esta informação é agora apresentada como resultado adicional à polaridade.

A polaridade geral de um comentário é apresentada de forma binária, positiva ou negativa, sendo calculada pelo somatório das polaridades presentes num comentário. A tabela 6.9 mostra a relação entre o valor obtido da polaridade e a transformação para uma representação binária.

Apesar de a polaridade binária identificar se um comentário é, na sua generalidade, positivo

Valor	Polaridade
<0	Negativa
== 0	Positiva
>0	Positiva

Tabela 6.9: Representação binária da polaridade.

ou negativo, a implementação do modelo permite obter informação mais detalhada acerca da polaridade obtida. Embora um comentário seja, por exemplo, positivo, com a polaridade binária não é possível distinguir entre outro comentário que seja também positivo. Esta distinção entre uma mesma polaridade pode ser dada pela magnitude da mesma – o quanto positivo ou negativo é um comentário.

A utilização, por si só, do valor do somatório das polaridades de um comentário, não permite representar a magnitude dado que não existe um valor mínimo e máximo de referência para comparação dos valores. Foi, portanto, utilizada uma escala de 1 a 5 onde 1 representa o menor valor de magnitude que corresponde a um comentário muito pouco positivo/negativo e 5 representa o maior valor de magnitude que corresponde a um comentário muito positivo/-negativo. Destaca-se a distinção entre a polaridade e a magnitude, um valor de magnitude não representa uma polaridade mas sim o quanto "forte" é uma polaridade. É, então, efetuada uma redução do valor do somatório das polaridades para uma escala de valores decimais entre 1 e 5 e é dada pela fórmula (que teve inspiração no trabalho efetuado em [23]):

$$Magnitude = 2 * (\frac{1}{2} * (\log_2(|\sum \text{pontuações}|))) + 3 \tag{6.1}$$

onde,

- $|\sum \text{pontuações}|$ é o valor absoluto do somatório das polaridades
- Se $\sum \text{pontuações} = 0.0$ então a magnitude é automaticamente igual a 1.0.
- $\frac{1}{2} * (\log_2(|\sum \text{pontuações}|))$ é limitado entre os valores -1 e 1.

Considere-se agora os comentários abaixo, do conjunto de dados, com respetivas representações em formato de lista de frases, que servem de exemplo para demonstrar toda a informação que é obtida/apresentada na aplicação do modelo sobre os mesmos:

C1: *"Poderia ser de mais fácil utilização os programas nem sempre são fáceis de utilizar"*

[['poderia', 'ser', 'mais', **-1.0**, 'utilização', 'programas', 'nem', 'sempre', 'são', **-1.0**, 'utilizar']]

C2: *"Máquina de excelente qualidade. Contudo a sua utilização é demasiado complexa para utilizadores principiantes"*

[['máquina', **2.0**, 'qualidade', '.'], ['contudo', 'utilização', 'é', 'demasiado', **-1.6**, 'utilizadores', 'principiantes']]

C3: *"Tal e qual conforme esperava, excelente qualidade, rápida entre funções e o som espectacular!"*

[['tal', 'qual', 'conforme', 'esperava', ','], [**2.0**, 'qualidade', ','], [**1.0**, 'funções', 'som', **1.0**, '!']]

A informação obtida pelo modelo para cada comentário pode ser consultada na tabela seguinte:

	C1	C2	C3
Contexto	Eletrodomésticos	Eletrónicos	Eletrónicos
Polaridade	Negativo	Positivo	Positivo
Magnitude	4.0	1.7	5.0
Aspetos positivos	[]	[qualidade]	[qualidade, funções, som]
Aspetos negativos	[utilização, programas]	[utilização]	[]

O modelo consegue prever corretamente toda a informação apresentada. O contexto e os aspetos tanto positivos como negativos são obtidos ao longo do processamento de cada comentário em processos específicos para esse efeito ou em processos cujo resultado serve de input para outros processos como é o caso da determinação do contexto, que é utilizado para a seleção da polaridade das palavras dependentes do contexto. Os aspetos são apresentados em listas nas quais uma lista vazia significa que não foram detetados aspetos explícitos do tipo em questão. A polaridade é binária, tal como já foi mostrado anteriormente. A análise dos comentários e das pontuações das polaridades neles contidos e corretamente identificadas (e com as várias transformações como modificadores de sentimentos, intensificadores e tonalidade) permite mostrar que o valor da magnitude atribuído à polaridade de cada comentário é o mais adequado. O primeiro comentário apenas tem informação negativa logo possui uma magnitude próxima do valor máximo (5), mas não o atingindo dado que a negatividade presente não é considerada extrema. O segundo comentário possui quase que um equilíbrio nas polaridades sendo, no entanto, a polaridade positiva ligeiramente mais acentuada que confere uma

polaridade positiva ao comentário e com magnitude abaixo da média da escala mostrando que apesar de positivo, não o é com muita intensidade. Por último, o terceiro comentário é composto exclusivamente por palavras positivas, uma de grande intensidade, que torna o comentário muito positivo, logo o valor da magnitude é igual a 5.

6.3 Avaliação do Modelo e Conclusões

No modelo supervisionado de aprendizagem automática foi utilizado o resultado da AUCROC como métrica de comparação e análise da qualidade do modelo. No modelo baseado em dicionários não é possível utilizar essa métrica para avaliação dado que é um modelo desenvolvido de forma manual cujos dados necessários para o cálculo da AUCROC não são possíveis obter. Uma primeira sensibilidade sobre os resultados é, então, obtida pela análise dos valores da matriz de confusão e métricas possíveis calcular a partir desses valores. Ainda no modelo supervisionado, os valores obtidos da matriz de confusão são referentes ao teste do modelo sobre o conjunto de dados de teste. No modelo baseado em dicionários não existe uma distinção entre um conjunto de dados de treino e de teste dado que o modelo não é treinado. De forma a manter uma correspondência de comparação dos resultados deste modelo com o de aprendizagem automática, nas figuras 6.9 e 6.10 e na tabela 6.10 apresentam-se os valores obtidos, tanto para todo o conjunto de dados como para apenas os dados de teste utilizados no modelo supervisionado, respetivamente.

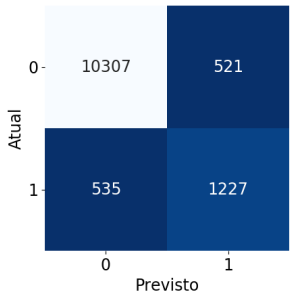


Figura 6.9: Matriz de confusão (todos os dados).

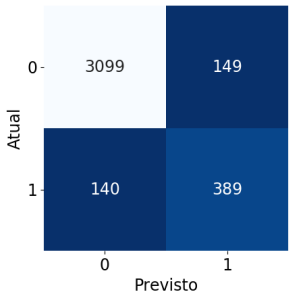


Figura 6.10: Matriz de confusão (dados de teste).

	Todos os dados	Dados de teste
Accuracy / Null_Accuracy	0.9161 / 0.86	0.9235 / 0.86
Precision	0.9507	0.9568
Sensitivity	0.9519	0.9541
Specificity	0.6964	0.7353

Tabela 6.10: *Accuracy*, *Precision*, *Sensitivity* e *Specificity* para todos os dados e para os dados de teste.

Os resultados obtidos são muito bons para um modelo baseado em dicionários onde apenas uma pequena percentagem de comentários são incorretamente classificados em relação à sua polaridade – quando aplicado sobre todo o conjunto de dados, apenas 1056 comentários possuem uma polaridade incorreta, o que representa aproximadamente 8.4% dos comentários, e quando aplicado apenas sobre o mesmo conjunto de dados utilizados para teste do modelo supervisionado, apenas 289 comentários são incorretamente classificados que representa aproximadamente 7.7% dos comentários.

Dado que a implementação deste modelo teve uma forte componente de análise/leitura de uma grande parte dos comentários do conjunto de dados, permitiu constatar a ocorrência de um problema comum a soluções de análise de sentimentos e relacionado com a anotação dos comentários com a respetiva polaridade – a subjetividade dos comentários. Os resultados são obtidos pela comparação da polaridade original dos comentários e da polaridade obtida pelo modelo. Contudo, a polaridade original está sujeita ao ponto de vista do autor que escreveu o comentário, que pode considerar que uns certos aspetos são mais relevantes do que outros mesmo sem os intensificar, atribuindo uma polaridade apenas de acordo com esses aspetos e excluindo os outros por completo. Considere-se o comentário abaixo do conjunto de dados representativo deste cenário:

C1: *"O ecrã é ótimo. A bateria dura muito pouco."*

A polaridade original, dada pelo autor do comentário é negativa. O comentário apresenta o ponto de vista do autor em relação a dois aspetos do produto e até contém alguma intensidade na forma como foi expresso pela utilização das palavras *ótimo* e *muito pouco* que quando inseridas no contexto representam duas polaridades opostas, uma positiva e outra negativa. Considere-se agora a representação do comentário em lista de frases quando aplicado o modelo sobre a mesma:

[['ecrã', 'é', **2.0**, '.'], ['bateria', 'dura', **-1.8**, '.']]

A polaridade obtida pelo modelo é positiva $(2.0 + (-1.8) = 0.2)$ e com baixa magnitude (pouco positiva) de acordo com as polaridades acima presentes no comentário, que estão

sujeitas à intensidade definida cujo processo foi anteriormente apresentado. O resultado do modelo é, então, incorreto dado que a polaridade original do mesmo é negativa. Isto deve-se ao facto de o autor apresentar dois pontos de vista com polaridades distintas, sem intensificar de forma relevante qual delas é verdadeiramente importante para ele, que neste caso é a duração da bateria. Este tipo de casos onde a determinação da polaridade está dependente da subjetividade do autor são muito difíceis de resolver, para um outro autor o aspeto ecrã poderia ser mais importante transformando a polaridade original do comentário como positiva que levava a que o modelo estivesse correto. Pode-se, assim, teorizar que o resultado do modelo pode ser considerado como correto quando o mesmo é a combinação da polaridade positiva com a baixa magnitude que seria também verdade para uma polaridade negativa com baixa magnitude caso a polaridade original fosse positiva. Para além de uma baixa percentagem de comentários com polaridade incorretamente definida, os resultados da matriz de confusão, tanto para todos os comentários como para apenas os de teste, têm um equilíbrio praticamente perfeito nas classes de erro (521 – 535; 149 – 140) que mostra que o modelo não é tendencioso para nenhuma das classes.

Apesar dos bons resultados, o modelo não é perfeito. Várias das transformações que constituem o modelo necessitam de ser melhoradas, de modo a abrangerem mais situações que ocorrem com menos frequência. De referir:

- Na solução de deteção de contexto teve de ser efetuada uma redução dos possíveis contextos para apenas dois, mais abrangentes, que leva a uma perda de precisão e por conseguinte a uma possível atribuição de uma polaridade errada a uma palavra de sentimento. A otimização desta solução para melhor reconhecimento dos vários contextos seria assim uma mais valia.

- A realização de uma revisão manual ao dicionário de sentimentos desenvolvido na qual foi corrigida a polaridade de várias palavras e expressões comuns, como já foi mostrado no respetivo capítulo. Contudo, a ocorrência de polaridades incorretamente atribuídas é algo que não pode ser colocado de parte sendo necessário uma análise ainda mais cuidada ao dicionário.

- A deteção de modificadores de sentimentos (*sentiment shifters*) consegue definir uma palavra como um modificador quando a mesma se encontra a uma distância de até 3 palavras da polaridade que modifica e que faz com que a associação de uma palavra modificadora a uma polaridade seja efetuada pelo princípio da proximidade entre eles. Contudo, existem casos onde a palavra modificadora de sentimentos se encontra ainda mais distante da polaridade, estes casos não são detetados pela solução atual. A otimização da solução de modo a reconhecer estes casos, possivelmente baseada em regras gramaticais ou até a utilização de aprendizagem automática, seria, assim, uma

mais valia dado a importância que esta transformação possui sobre a polaridade final.

- A aplicação de intensificadores de sentimentos é baseada num conjunto de palavras e expressões de intensidade manualmente desenvolvido e de acordo com um conjunto de regras de aplicação. Este conjunto não possui, contudo, todas as possíveis palavras de intensidade que leva a uma futura necessidade de identificar mais palavras e expressões deste tipo. Para além disto, embora o conjunto de regras de aplicação funcione bem, existem sempre exceções que devem ser consideradas para melhoria da solução.

- A tonalidade de sentimentos tenta dar relevância às polaridades dos aspetos que o autor mostrou serem mais relevantes para ele. Como nas transformações anteriores, esta também é baseada num conjunto manualmente desenvolvido de conjunções adversativas e regras de aplicação. A atualização deste conjunto com novas conjunções e aperfeiçoamento das regras para que sejam menos restritas e consigam associar mais corretamente uma conjunção adversativa ao conjunto de polaridades que anula, evitando anular polaridades por ela não afetadas ou, de facto, anular polaridades que devam ser anuladas mas que não o são, seria o próximo ponto de otimização.

- A deteção dos aspetos positivos e negativos explicitamente presentes nos comentários é o processo de maior complexidade. Está dependente da correta marcação das palavras com a respetiva parte do discurso dado que os nomes/substantivos são utilizados para representar aspetos. Uma melhoria no *POS Tagger* desenvolvido trás assim grandes benefícios para esta transformação, fazendo com que palavras deixem de ser incorretamente marcadas como nomes/substantivos. Para além disto, existem ainda aspetos que são implícitos, não representados por nomes/substantivos, que a solução não se preocupa em detetar e que pode ser uma forma de otimização. A deteção dos aspetos é também baseada na proximidade com a polaridade que os define e de acordo com um conjunto de regras que em muitos casos não são aplicáveis levando ao não reconhecimento de um aspeto que, muito provavelmente, se encontra distante da polaridade e é referenciado por um pronome demonstrativo. A deteção destes casos seria também o próximo passo a desenvolver.

7 O Modelo Híbrido

Uma vez desenvolvidas as duas soluções de análise de sentimentos em textos, uma baseada num modelo supervisionado de aprendizagem automática e outra baseada em dicionários e léxicos, foi desenvolvido um modelo híbrido. O objetivo é, portanto, tirar partido dos pontos fortes de cada um dos modelos desenvolvidos e unifica-los apenas num, aumentando potencialmente a qualidade dos resultados. Assim, este modelo híbrido fará uso dos resultados dos dois modelos anteriores e agrega-os de modo a obter uma nova polaridade para cada comentário, com os pesos ponderados dos resultados dos dois modelos. A figura 7.1 mostra o esquema geral deste modelo, bastante simples, cujo processamento dos comentários é efetuado pelos modelos anteriormente desenvolvidos e de igual modo como já foi exposto.

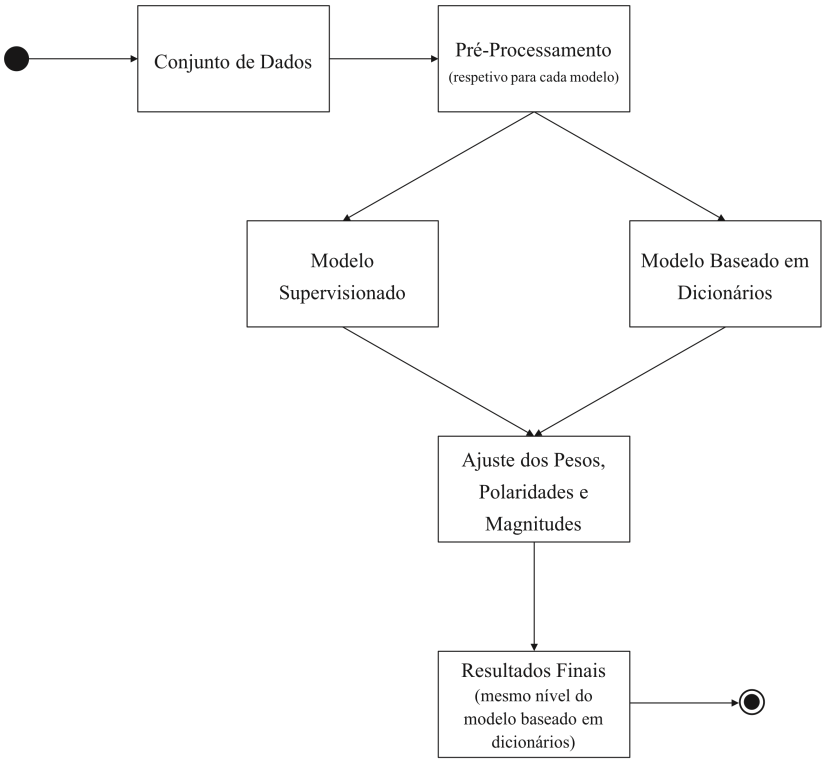


Figura 7.1: Arquitetura modelo híbrido.

7.1 Metodologia

Por pré-definição, o modelo híbrido adota inicialmente o mesmo conjunto de resultados obtidos pelo modelo baseado em dicionários. Dado que este último é um modelo ao nível do aspeto, esta adoção dos seus resultados proporciona ao modelo híbrido o mesmo nível. O processo associado com o modelo híbrido é, depois, executado da seguinte maneira:

- Comparação da polaridade obtida nos dois modelos.
- Se os dois modelos estão de acordo com a polaridade então é mantida toda a informação resultante do modelo baseado em dicionários, dado que é o que apresenta uma maior quantidade de informação.
- Para um comentário cuja polaridade obtida pelos dois modelos não seja a mesma então a nova polaridade do comentário é calculada tendo em conta os pesos do resultado de cada modelo pela aplicação da fórmula:

$$Polaridade = P_{aprendizagem} + P_{\text{léxico}} \tag{7.1}$$

onde,

$P_{\text{léxico}} = \sum pontuações$

$P_{aprendizagem} = (\pm)(2 * (\log_2(probabilidade)) + 3)$

O peso do modelo baseado em dicionários ($P_{\text{léxico}}$) é dado pelo somatório das polaridades atribuídas a cada palavra de sentimento reconhecida, tal como já foi exemplificado várias vezes ao longo do respetivo capítulo.

O peso do modelo de aprendizagem automática é ligeiramente mais complexo. Como é um modelo de classificação, a cada possível classe (neste caso apenas duas: 0 e 1) está associado um peso representado por uma probabilidade, no qual a soma das probabilidades das duas classes é igual a 1. A classe que tiver a maior probabilidade é, portanto, a que é considerada como resultado. Na fórmula acima, a variável *probabilidade* é, então, o valor da probabilidade obtida na classe resultado. Os valores da probabilidade encontram-se entre [0.5, 1.0]. Esta probabilidade é transformada pelo resto da fórmula num intervalo de valores próximo dos valores médios obtidos pelo modelo baseado em dicionários, possibilitando assim a associação dos dois modelos. A escolha do sinal é efetuada de acordo com a classe resultado: se positiva, então o sinal é positivo, senão o sinal é negativo; tornando assim os valores possíveis de $P_{aprendizagem}$ dentro do intervalo $(\pm)$[1.0, 3.0].

No modelo de aprendizagem automática é ainda efetuada uma posterior alteração ao valor de $P_{aprendizagem}$ dada por:

$$P_{aprendizagem} = P_{aprendizagem} * (0.5)$$

Este ajuste do peso de $P_{aprendizagem}$ para metade do valor originalmente obtido é efetuado apenas quando a classe resultado é positiva. Na análise dos resultados do melhor modelo de aprendizagem foi discutido a tendência do modelo para a classe positiva. Como o modelo baseado em dicionários possui classes equilibradas houve a necessidade de diminuir o peso que a classe positiva possui sobre o modelo híbrido. A aplicação de várias percentagens de redução mostrou que a redução em metade do peso é a que obtém os melhores resultados.

A nova polaridade é, desta forma, obtida pela soma dos pesos das polaridades dos dois modelos, que são garantidamente um positivo e outro negativo. Assim, por exemplo, caso o modelo de aprendizagem automática tenha um resultado com baixo peso positivo e o modelo baseado em dicionários tenha um peso médio negativo, a nova polaridade vai ser igual a um baixo negativo. Como foi aplicado até aqui, um valor inferior a 0 resulta numa polaridade negativa e um valor igual ou superior a 0 resulta numa polaridade positiva. Este método utilizado para cálculo da nova polaridade, resultante da ponderação dos dois modelos, permite ainda facilmente determinar a magnitude desta nova polaridade pela utilização da fórmula da magnitude já apresentada no modelo baseado em dicionários agora com o valor da polaridade híbrida. Assim, este modelo apresenta a mesma quantidade de informação que é apresentada pelo modelo baseado em dicionários, onde a informação do contexto e dos aspetos positivos e negativos são calculados por ele e podem ser aqui utilizados, dado que a sua determinação não está dependente da polaridade. A polaridade final e magnitude são dadas também pelo modelo baseado em dicionários, caso os dois modelos (supervisionado e léxico) estejam de acordo com a polaridade do comentário, ou pela aplicação híbrida dos resultados caso contrário.

7.2 Análise e Comparação dos Resultados

Uma vez que o modelo híbrido utiliza o modelo supervisionado de aprendizagem faz sentido analisar os resultados apenas sobre os dados de teste que também foram utilizados para avaliar este último. De modo a poder comparar melhor os resultados dos três modelos principais desenvolvidos, veja-se as figuras 7.2, 7.3 e 7.4, que apresentam as matrizes de confusão de cada um dos modelos e a tabela 7.1, que mostra informação de outras métricas possivelmente relevantes.

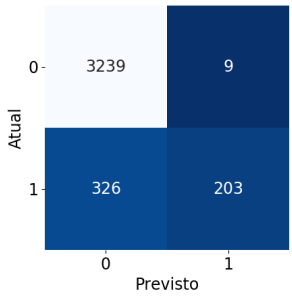


Figura 7.2: Matriz de confusão modelo aprendizagem automática (dados teste).

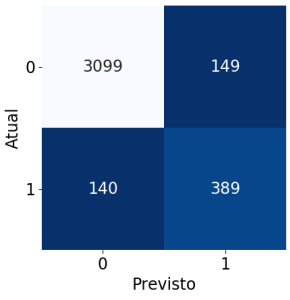


Figura 7.3: Matriz de confusão modelo dicionários (dados teste).

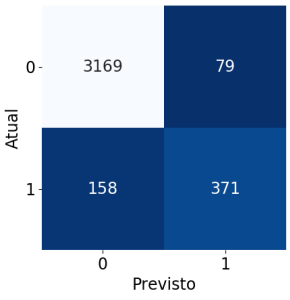


Figura 7.4: Matriz de confusão modelo híbrido (dados teste).

	Modelo aprendizagem automática	Modelo baseado em dicionários	Modelo híbrido
Accuracy / Null_Accuracy	0.9113 / 0.86	0.9235 / 0.86	0.9373 / 0.86
Precision	0.9086	0.9568	0.9525
Sensitivity	0.9972	0.9541	0.9757
Specificity	0.3837	0.7353	0.7013

Tabela 7.1: *Accuracy*, *Precision*, *Sensitivity* e *Specificity* para todos os modelos (dados de teste).

No modelo supervisionado de aprendizagem automática foi utilizado o valor de AUCROC para o avaliar e no modelo baseado em dicionários foi utilizado, maioritariamente, a quantidade de comentários corretamente e incorretamente classificados, dado que AUCROC é uma métrica possível de obter apenas em modelos de aprendizagem automática que leva a que o modelo híbrido também não possa ser avaliado por esta métrica. Neste sentido, foi utilizado dois métodos de avaliação do modelo, nomeadamente o:

- Somatório do número de comentários incorretamente classificados.
- *MCC – Matthews Correlation Coefficient.*

A importância do primeiro método é clara: quanto menos classificações incorretas ocorrerem potencialmente melhor será o modelo. Contudo, isto não basta para garantir que um modelo é, de facto, melhor que outro, quando as classes não são balanceadas (como já foi demonstrado na secção 5.2 do capítulo 5). O *Matthews Correlation Coefficient* (MCC) é, então, utilizado dado que é uma métrica capaz de lidar com classes não balanceadas e assim apropriada para comparar os modelos. A figura 7.5 mostra os valores obtidos nesta métrica para cada um dos modelos.

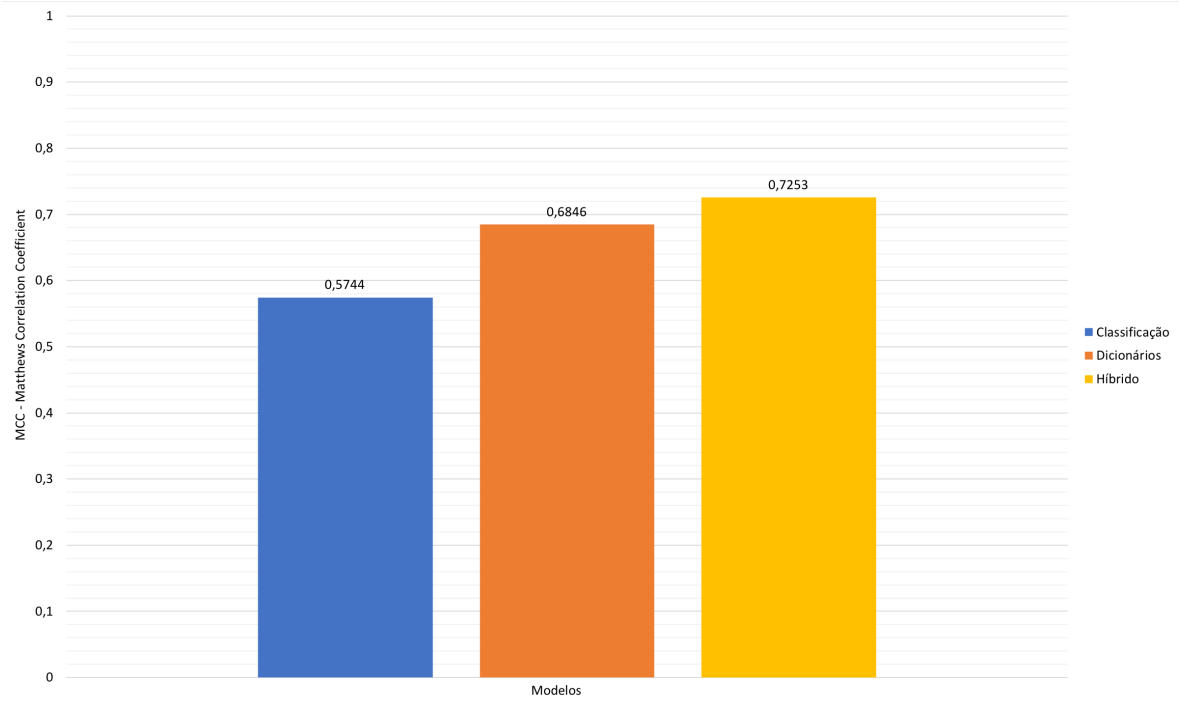


Figura 7.5: MCC – Matthews Correlation Coefficient para os modelos.

O MCC é uma métrica que utiliza a ponderação de todos os valores da matriz de confusão, ou seja, tanto os verdadeiros positivos e negativos como os falsos positivos e negativos. O MCCé, assim, uma métrica balanceada e capaz de ser aplicada sobre qualquer conjunto de dados, mesmo quando as classes são muito desbalanceadas. O resultado desta métrica é um valor entre [-1, +1], no qual um valor igual a +1 representa um modelo perfeito com previsões perfeitas e 0 representa um modelo cujas previsões são simplesmente aleatórias caracterizando o modelo como muito mau dada a aleatoriedade dos resultados. Os valores negativos representam uma previsão reversa com o mesmo peso de valores (-1 representa uma previsão perfeita/modelo perfeito mas cujos resultados estão revertidos nas classes). Na figura 7.5 é necessário ter atenção a escala escolhida para os valores de MCC, embora a gama de valores possíveis seja entre -1 e +1, a figura utiliza uma escala entre 0 e +1 dado que não foram obtidos valores negativos e estes representam apenas uma previsão reversa, sendo assim mais importante o resultado dos valores entre a escala escolhida.

Os valores de MCC obtidos nos modelos foram bons e de forma crescente de modelo para modelo (figura 7.5), sendo que o modelo híbrido possui o maior valor, indo assim de encontro com as expectativas da sua implementação como forma de unificar e melhorar os dois modelos anteriores. A análise da matriz de confusão do modelo híbrido (figura 7.4) mostra uma redução significativa do número de comentários incorretamente classificados em comparação com os modelos anteriores e uma redução da tendência para a classe positiva que existia no modelo de

classificação (a redução para metade do peso que a classe positiva do modelo de classificação tem sobre a ponderação dos dois modelos foi um fator de grande importância para conseguir isto) que em conjunto com o valor de MCC obtido, permite defender que o modelo híbrido é, de facto, o melhor modelo dos três. Note-se ainda que, embora os valores obtidos de MCC nos modelos sejam matematicamente próximos, como a escala possível de valores é pequena, uma diferença de 0.1 representa já um grande impacto na qualidade do modelo.

8 Conclusões e Trabalho Futuro

8.1 Conclusões Finais

Conseguir aproximar um computador às capacidades do ser humano é algo que está cada vez mais presente na nossa sociedade à medida que a tecnologia evolui e novas soluções e métodos de processamento emergem. O processamento de linguagem natural é um tema que desde há muito tempo tem sido alvo de grande análise e que nos últimos anos tem sido utilizado em grande escala com a implementação de assistentes virtuais nos mais variados domínios de aplicação. Do grande leque de possíveis aplicações do processamento de linguagem natural, a análise dos sentimentos presentes em textos é uma área de trabalho de grande dificuldade, dada a grande subjetividade associada aos sentimentos e às diversas dificuldades inerentes à sua identificação e classificação em textos, tal como já foi demonstrado ao longo deste trabalho de dissertação.

O pré-processamento dos dados é uma fase de extrema importância em qualquer processo de análise de dados. Como tal deve ser efetuado de forma ponderada, dadas as suas possíveis consequências que tanto podem ser boas como más a longo prazo. De facto, esta fase, de uma forma geral, costuma e deve ocupar tanto tempo como o tempo necessário para desenvolvimento de um modelo. Um correto pré-processamento permite não só ter um conhecimento absoluto do conjunto de dados e da área (ou do problema) de análise como é responsável por efetuar todas as transformações como remoção de dados não relevantes, normalização, acréscimo de informação tanto manual como automática, entre outros. O pré-processamento de texto é uma tarefa com grau de dificuldade especialmente elevado dadas as grandes possibilidades de exploração dos dados. Como foi mostrado no pré-processamento efetuado neste trabalho de dissertação, cada transformação pode assentar sobre o resultado de uma anterior, em que podem ser analisados os dados a cada intervalo de transformações. Esta característica foi determinante para conseguir perceber a forma como os comentários são comummente escritos e, por conseguinte, fazer a sua limpeza efetuando, por exemplo, a correção de erros ortográficos frequentes, a remoção de *stop words* e de outras palavras não relevantes obtidas pela análise do seu tamanho, bem como a sua frequência e co-ocorrência com outras palavras.

Vários modelos de aprendizagem automática supervisionada foram desenvolvidos com vários classificadores. Em qualquer problema de aprendizagem automática é essencial que mais do que um classificador seja utilizado dado que, dependendo do problema e dos dados utilizados. Um dado classificador não é garantidamente sempre o mais adequado para um mesmo problema. Para além disto, no processamento de texto, sempre que possível devem ser efetuados vários modelos de forma a explorar as possíveis representações do texto. Nenhum classificador é capaz de processar texto no seu estado normal, o que leva à necessidade de transformação em vetores representativos do mesmo. Várias representações dos comentários em respetivos vetores foram desenvolvidas. Para tal foram utilizadas transformações típicas da representação de texto para uso em modelos de aprendizagem automática, em que cada modelo tinha o objetivo de tentar melhorar o modelo anterior. Algumas dessas representações passaram pela transformação dos comentários numa matriz de vetores de presença de palavras, frequência de palavras, conjuntos de n-gramas, redução do vocabulário para apenas certas partes do discurso e até utilização de métodos mais avançados como *Word2Vec*. Surpreendentemente, o melhor modelo conseguido é o que faz uso de uma das transformações mais simples – representação em frequência de palavras sem acentos e sem sufixos com classificador de *random forests*. Todos os modelos apresentam resultados próximos uns dos outros. O melhor modelo apresenta uma tendência notável para a classe positiva devido ao não-balanceamento do conjunto de dados que pode ser considerado o próximo ponto de partida para melhoria do modelo.

O modelo baseado em dicionários permite uma maior liberdade de implementação que levou a que fosse possível aumentar substancialmente a quantidade de informação final apresentada como resultado. Deteção do contexto, forma como se deteta e lida com modificadores de sentimentos, aplicação de intensidade e tonalidade aos sentimentos, deteção de aspetos positivos e negativos presentes em cada comentário e ainda calculo da magnitude de uma polaridade é o conjunto de informação que é conseguida obter com este modelo. Utiliza por base o dicionário de sentimentos desenvolvido e todo um conjunto de pequenos léxicos e até aprendizagem automática são utilizados para obtenção desta informação. Praticamente todos os métodos desenvolvidos têm por base a noção de proximidade entre uma certa palavra e a polaridade que a representa. Esta forma de processamento mostrou-se surpreendentemente eficaz para o domínio de análise, tendo este modelo superado os resultados do modelo supervisionado de aprendizagem automática. Contudo, existe ainda um grande espaço de melhoria do modelo.

Como foi visto neste trabalho, o modelo híbrido (capítulo 7) é um modelo simples que visa tirar partido dos pontos fortes dos dois modelos anteriormente desenvolvidos. Toda a complexidade é efetuada nos respetivos modelos sendo da responsabilidade do modelo híbrido agregar os resultados de cada um. A tendência do modelo supervisionado para a classe positiva foi minimizada pela redução do peso numa previsão desse tipo por esse modelo. O tipo e a quantidade de resultados (nível de análise) do modelo híbrido é o mesmo do modelo baseado

em dicionários e os resultados conseguem ser os melhores de entre os modelos servindo de argumento para sustentar que uma solução de análise de sentimentos não deve restringir-se a um modelo mas sim utilizar vários modelos para os vários processos constituintes.

Qualquer um dos modelos desenvolvidos e recursos que utilizam, quer sejam manuais ou modelos de aprendizagem automática (como o detetor de contexto), foram gravados de forma persistente e estão, assim, prontos a serem inseridos em qualquer outra solução de análise de sentimentos como por exemplo uma interface Web. Todos os ficheiros e léxicos construídos manualmente estão também em formato tabulado (na sua maioria são ficheiros *.csv*), o que facilita uma eventual portabilidade do conteúdo para uma base de dados, algo indispensável numa qualquer solução informática.

8.2 Trabalhos Futuros

Na nossa opinião, a qualidade dos resultados obtidos foi surpreendentemente boa, dada a dificuldade característica desta área de análise na obtenção de resultados do tipo *estado da arte*. Os resultados obtidos não podem ainda ser assim classificados, mas a implementação de um possível conjunto de melhorias pode levar à obtenção de resultados mais próximos dessa meta, como:

- Redução da tendência para a classe positiva do modelo supervisionado. O equilíbrio na quantidade de classificações incorretas é essencial para que o modelo seja considerado aceitável ou mesmo bom. Dado que a classe positiva é a mais tendenciosa, isto significa que o modelo tem dificuldade em distinguir os comentários negativos, potencialmente devido à não deteção de modificadores de sentimentos (*sentiment shifters*), que são os maiores responsáveis por inverter a polaridade de uma frase positiva. Uma melhor identificação de modificadores de sentimentos é, pois, essencial e pode ser conseguido tirando partido do mesmo método utilizado no modelo baseado em dicionários. Sempre que um modificador de sentimentos seja detetado, pode-se acrescentar uma *tag* a todas as palavras por ele afetadas (o que leva naturalmente a um aumento do vocabulário) que para o modelo supervisionado é mais fácil distinguir ou de associar à negatividade presente.

- Apesar de terem sido desenvolvidos vários modelos de aprendizagem automática muitos outros podiam ser também concebidos e implementados. As hipóteses são muitas, mas há alguns que claramente são os próximos candidatos. Tal é o caso de uma solução de aprendizagem automática em redes neuronais ou de *deep learning*. O desenvolvimento de um modelo deste tipo tem grande interesse não só pela potencialidade de melhoria

dos resultados mas também em efeitos comparativos com o melhor modelo supervisionado conseguido. A utilização de redes neuronais/*deep learning* não é garantidamente melhor do que os métodos de aprendizagem mais tradicionais, vários fatores podem influenciar negativamente a capacidade de aprendizagem de modelos desse tipo como a dimensão e qualidade do conjunto de dados.

- Revisão mais profunda do dicionário de sentimentos desenvolvido. A ocorrência de polaridades incorretas não é algo que possa ser descartado. Enriquecimento dos léxicos criados como o léxico de expressões comuns que devem ser consideradas como apenas uma palavra. A aplicação desta transformação sobre os modelos de aprendizagem automática pode também trazer muitas melhorias.

- Elaboração de um método de deteção de contextos mais eficiente, mais sensível à presença de apenas uma palavra de entre muitas que é a que de facto define o contexto do texto de análise.

- Muitos dos processos desenvolvidos no modelo baseado em dicionários fazem uso da noção de proximidade entre uma polaridade e as palavras por ela afetadas ou que a afetam. Esta proximidade é um valor relativamente fixo, definido na implementação do processo que pode não ser suficiente quando essa proximidade é mais distante. A implementação de uma solução capaz de identificar estes casos seria então uma mais valia.

- Como não poderia deixar de ser, a implementação de um modelo híbrido que tire partido das melhorias dos novos modelos.

Estes são apenas algumas das linhas de trabalho que poderíamos adotar e desenvolver para melhorar a solução desenvolvida. No capítulo 2 foram apresentados os problemas e as dificuldades mais comuns em análise de sentimentos em textos, problemas estes que nos modelos desenvolvidos se tentou minimizar o seu efeito.

Bibliografia

[1] Projecto floresta sintá(c)tica. https://www.linguateca.pt/floresta/.

[2] Snowballstemmer. http://snowball.tartarus.org/.

[3] Mariana S. C. Almeida, Claudia Pinto, Helena Figueira, Pedro Mendes, and André F. T. Martins. Aligning opinions: Cross-lingual opinion mining with dependencies, 2015.

[4] Orestes Appel, Francisco Chiclana, Jenny Carter, and Hamido Fujita. A hybrid approach to the sentiment analysis problem at the sentence level. *Knowledge-Based Systems*, 108:110–124, September 2016.

[5] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE, 2010.

[6] Carmen Banea, Rada Mihalcea, and Janyce Wiebe. Sense-level subjectivity in a multilingual setting. *Computer Speech & Language*, 28(1):7–19, January 2014.

[7] Bird, Steven, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 1ª edition, 2009.

[8] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. 2010.

[9] Felipe Bravo-Marquez, Marcelo Mendoza, and Barbara Poblete. Meta-level sentiment models for big social data analysis. *Knowledge-Based Systems*, 69:86–99, October 2014.

[10] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 107–116. Association for Computational Linguistics, 2010.

[11] Xiaowen Ding, Bing Liu, and Philip S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data*

mining, pages 231–240. ACM, 2008.

[12] Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. Identifying sarcasm in Twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*, pages 581–586. Association for Computational Linguistics, 2011.

[13] Vasileios Hatzivassiloglou and Janyce Wiebe. Effects of Adjective Orientation and Gradability on Sentence Subjectivity. 2000.

[14] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.

[15] IBM. CRISP-DM Guide, 2002. OCLC: 728715922.

[16] Jaap Kamps, Maarten Marx, Robert Mokken, and Maarten Rijke. Using WordNet to Measure Semantic Orientations of Adjectives. 2004.

[17] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics, 2004.

[18] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[19] Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: analyzing and comparing opinions on the web. *Proceedings of the 14th international conference on World Wide Web*, pages 342–351, 2005.

[20] Isa Maks and Piek Vossen. A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems*, 53(4), November 2012.

[21] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, December 2014.

[22] M. Dolores Molina-González, Eugenio Martínez-Cámara, María-Teresa Martín-Valdivia, and José M. Perea-Ortega. Semantic orientation for polarity classification in Spanish reviews. *Expert Systems with Applications*, 40(18):7250–7257, December 2013.

[23] Andrius Mudinas, Dell Zhang, and Mark Levene. Combining lexicon and learning based approaches for concept-level sentiment analysis. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 5. ACM, 2012.

[24] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[25] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

[26] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. 1(2):91–231, 2008.

[27] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86, 2002.

[28] Kumar Ravi and Vadlamani Ravi. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46, November 2015.

[29] John Rothfels and Julie Tibshirani. Unsupervised sentiment classification of English movie reviews using automatic selection of positive and negative sentiment items. *CS224N-Final Project*, 43(2):52–56, 2010.

[30] Beatrice Santorini. Part of Speech Tagging Guidelines for the Penn Treebank Project. 1990.

[31] Sunita Sarawagi. Information extraction. *Foundations and Trends® in Databases*, 1(3):261–377, 2008.

[32] Mário J. Silva, Paula Carvalho, Carlos Costa, and Luís Sarmento. Automatic expansion of a social judgment lexicon for sentiment analysis technical report, December 2010.

[33] Mikalai Tsytsarau and Themis Palpanas. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3):478–514, May 2012.

[34] Peter D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

[35] Xiaojun Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-volume 1*, pages 235–243. Association for Computational Linguistics, 2009.

[36] Suge Wang, Deyu Li, Xiaolei Song, Yingjie Wei, and Hongxia Li. A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. *Expert Systems with Applications*, 38(7):8696–8702, July 2011.

[37] Janyce Wiebe and Ellen Riloff. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In *CICLing*, volume 5, pages 486–497. Springer, 2005.