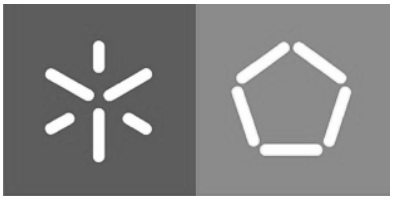


Universidade do Minho
Escola de Engenharia
Departamento de Informática

Hugo Alexandre Machado Fernandes

Mineração de Opiniões e de Sentimentos para o Estabelecimento de Índices de Satisfação

Novembro de 2018



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Hugo Alexandre Machado Fernandes

Mineração de Opiniões e de Sentimentos para o Estabelecimento de Índices de Satisfação

Dissertação de Mestrado
Mestrado em Engenharia Informática

Trabalho realizado sob a orientação de

Professor Doutor Orlando Manuel de Oliveira Belo

Novembro de 2018

Anexo 1

Declaração relativa ao depósito de dissertações e teses no RepositóriUM

Nome: Hugo Alexandre Machado Fernandes

Correio eletrónico: halexandremfernandes@gmail.com Tel./Telemóvel: 914736241

Número do Cartão de Cidadão/Bilhete de Identidade: 14582907

Título(s) da dissertação de mestrado ☑ / tese de doutoramento ☐: Mineração de Opiniões e Sentimentos para o Estabelecimento de Índices de Satisfação

Orientador(es): Orlando Manuel de Oliveira Belo

Data de conclusão: 23/10/2018

Mestrado e área de especialização/Doutoramento e especialidade: Mestrado em Engenharia Informática - Especializações em Business Intelligence e Engenharia de Aplicações

Declaro que concedo à Universidade do Minho o direito não-exclusivo e irrevogável de arquivar, reproduzir, comunicar e/ou distribuir através do seu repositório institucional, nas condições abaixo indicadas, a versão final da minha dissertação/tese em suporte digital, aprovada após a realização das provas de defesa pública e, quando for caso disso, após confirmação pelo(s) orientador(es) e homologação pelo presidente do júri[1] da introdução das alterações solicitadas.

Declaro que autorizo a Universidade do Minho a arquivar mais de uma cópia da dissertação ou tese e a, sem alterar o seu conteúdo, convertê-la para qualquer formato de ficheiro, meio ou suporte, para efeitos de preservação e acesso

Declaro que a dissertação ou tese agora entregue é um trabalho original e que, contendo material do qual não detenho direitos de autor, obtive autorização prévia do detentor dos referidos direitos para conceder à Universidade do Minho os termos requeridos por esta licença.

Declaro também que a entrega do documento não infringe, tanto quanto me é possível saber, os direitos de qualquer outra pessoa ou entidade.

Se o documento entregue é baseado em trabalho financiado ou apoiado por organismo/financiador que não a Universidade do Minho, declaro que cumpri quaisquer obrigações exigidas pelo respetivo contrato ou acordo.

Retenho todos os direitos de autor relativos à dissertação ou tese e o direito de a usar em trabalhos futuros, como artigos ou livros.

Concordo que a minha tese ou dissertação seja colocada no repositório da Universidade do Minho com o seguinte estatuto (assinale apenas um dos quatro estatutos):

1. ☑ Disponibilização imediata do conjunto do trabalho para acesso mundial
2. ☐ Disponibilização do conjunto do trabalho para acesso exclusivo na Universidade do Minho durante o período de ☐ 1 ano, ☐ 2 anos ou ☐ 3 anos, sendo que após o tempo assinalado autorizo o acesso mundial.
3. ☐ Disponibilização do conjunto do trabalho para acesso exclusivo na Universidade do Minho.
4. ☐ Outro (se assinalar este estatuto, a declaração deve ser obrigatoriamente acompanhada de uma exposição que explicite a necessidade de um estatuto de excecionalidade)

Braga/Guimarães, 23 / 10 / 2018

Assinatura: Hugo Alexandre Machado Fernandes

[1] *Vide* artigo 124.º Anexo ao Despacho RT-41/2014.

Agradecimentos

O desenvolvimento deste trabalho apenas foi possível com a ajuda, compreensão e disponibilidade de várias pessoas e entidades. Um conjunto de pessoas que me proporcionou e fez com que eu pudesse desenvolver o trabalho nas melhores condições possíveis.

Agradeço, pois, ao Professor Doutor Orlando Belo, pela oportunidade que me deu de trabalhar consigo nesta etapa académica, auxiliando-me para que pudesse atingir o meu objetivo. Agradeço, profundamente, toda a disponibilidade, compreensão e sabedoria para me ensinar e aconselhar em todas as ocasiões do desenvolvimento deste trabalho.

Aos meus Pais, à minha Irmã e à minha namorada, o meu mais profundo obrigado por me acompanharem e me proporcionarem, também, as condições necessárias para atingir mais um marco importante na minha vida académica, auxiliando-me na concretização de mais um objetivo a que me propus.

Agradeço também à DigitalSign – Certificadora Digital S.A., mais especificamente ao Dr. Fernando Moreira e ao meu chefe Serafim Gomes, que desde a minha entrada na Empresa me proporcionaram as melhores condições e a disponibilidade necessária para que pudesse dedicar parte do meu tempo ao desenvolvimento deste trabalho.

Resumo

Mineração de Opiniões e de Sentimentos para o Estabelecimento de Índices de Satisfação

Nunca como hoje as opiniões e os sentimentos de cada ser humano desempenharam um papel tão fundamental no quotidiano da sociedade.

A mineração de opiniões e de sentimentos providencia um grau de conhecimento bastante bom sobre o nível de satisfação que um dado negócio, um acontecimento, ou uma decisão estratégica pode gerar. Estar na posse de um índice de satisfação sobre um determinado bem, serviço, ou acontecimento é dispor de condições que nos permitem alcançar situações de sucesso que podem ser usadas para melhorar o planeamento ou a estruturação de uma decisão ou mesmo de um negócio, face ao feedback que usualmente provocam.

O conhecimento recolhido por este tipo de processo de mineração e, consequentemente, acumulado sobre a experiência que um certo cliente detém num determinado negócio, ou sobre a opinião (ou sentimento) que um certo indivíduo possui relativamente a um acontecimento de grande relevância político ou social permite definir novas estratégias de abordagem em determinados mercados. Neste trabalho de dissertação realizámos a implementação de um sistema de mineração de opiniões e de sentimentos tendo como finalidade de definir e estabelecer um dado índice de satisfação, de forma a que seja possível angariar conhecimento útil sobre aspetos que

potenciem a geração de opiniões e de sentimentos, potenciando novas oportunidades de negócio.

Palavras-Chave: Mineração de Opiniões, Mineração de Sentimentos, Índices de Satisfação, *Profiling*, Análise de Opiniões, Análise de Sentimentos, Categorização de Opiniões, Categorização de Sentimentos, *Feedback* do Consumidor, Mineração de *Críticas*, Métricas de Avaliação de Performance.

Abstract

Satisfaction Levels Establishment through Opinion and Sentiment Mining

Never like today, opinions and feelings of each human being, play a fundamental role on society's quotidian. Attending to the development state at what we arrived on the informations technology level, the individual opinion assumes an importance that erstwhile never had.

Knowing what an individual thinks, feels, and desires, and consequently, decides, on the most variety of all life dimensions, is fundamental in a global context. Knowing these feelings that could influence, directly and decisively, the most various behaves, is, on these days, of capital importance on the economics plan.

With the right available tools, we can increase the knowledge about behaves that could and should contribute to the improvement of economic structures performance.

Opinion and feeling mining provides a knowledge level about the satisfaction level that a business, an event, or a strategical decision generates.

Being on possession of a satisfaction level about a certain good, service, or event is arranging the conditions that aim to the success and that could be used to better plan a business or a decision, against the feedback that these ones provoke.

The collected knowledge, and, consequently, accumulated about the experience that a certain client has on a determined business, or about the opinion/feeling that a certain individual possesses relatively to a certain event of

enormous relevance at the political and social levels, enable the acquirement of specific strategies of approach on determined markets.

On the present dissertation we propose the implementation of an opinion and feeling mining system, generating like that the establishment of satisfaction indexes, so that we could obtain knowledge about the various aspects on cause, that generate the opinions and feelings, enabling by that way, the act of potentiate possible opportunities of business/businesses.

Keywords: Opinion Mining, Sentiment Mining, Satisfaction Indexes, *Profiling*, Opinion Analysis, Sentiment Analysis, Opinion Categorization, Sentiment Categorization, Consumer's *Feedback*, Reviews Mining, Performance Evaluation Metrics.

Índice

1 Introdução 1
1.1 Contextualização 1
1.2 Motivação e Objetivos 4
1.3 Estrutura do documento 13
2 Mineração de Opiniões e Sentimentos 17
2.1 Linha cronológica 23
2.1.1 A Interpretação de Texto (1979-1995) 24
2.1.2 Anotação de Opiniões de Baixo-Nível (1997-2003) 26
2.1.3 Diferença entre Subjetividade e Objetividade (2003-2005) 27
2.1.4 Aplicações Web de Mineração de Dados (2006-2007) 28
2.1.5 Recursos Léxicos (2008-2011) 29
3 Análise de Opiniões e Sentimentos 31
3.1 Níveis de Análise 36
3.2 Léxico Sentimental 39
3.3 Opiniões e Sentimentos 46
3.4 Polaridade dos dados 49
3.5 Classificação de Sentimentos 55
3.5.1 Classificação de Sentimentos Baseada em Aprendizagem supervisionada 57
3.5.2 Classificação de Sentimentos Baseada em Aprendizagem não supervisionada 61
3.6 Índices de Satisfação 66

4 Um Caso de Estudo 71
4.1 Contextualização da Aplicação 71
4.2 A Motivação para a Implementação da Aplicação 74
4.3 Objetivos da Aplicação 77
4.4 A Implementação 79
4.5 A Extração dos Dados 80
4.6 O Pré-processamento dos Dados 82
4.7 A Análise dos Dados Recolhidos 84
4.8 Disponibilização dos Resultados 87
4.9 Comparação de Resultados 96
4.9.1 O Tema "Brexit" 99
4.9.2 O Tema "Facebook" 101
4.9.3 O Tema "Corruption" 104
4.9.4 O Tema "Ryanair" 106
4.9.5 O Tema "McDonald's" 108
4.9.6 Conclusão dos Resultados dos Modelos 110
5 Conclusões e Trabalho Futuro 113
5.1 Conclusões 113
5.2 Trabalho Futuro 116
6 Bibliografia 119

Índice de Figuras

Figura 1: Processo de Sumariação baseado na estrutura dos tópicos. 7
Figura 2: Resultado da sumariação para o conjunto de críticas sobre o Nokia 6610. 8
Figura 3: Aplicação SentBuk.....9
Figura 4: Exemplo de crítica de ecommerce com recurso a emojis 42
Figura 5: Exemplo de crítica de cinema com recurso a emojis 43
Figura 6: Processo de análise de sentimentos em críticas de produtos 51
Figura 7: Exemplo de sumário de opiniões baseadas em características de um produto. 53
Figura 8: Visualização do sumário de opiniões baseadas em características do produto. 53
Figura 9: Comparação visual entre as opiniões sobre dois produtos. 54
Figura 10: Técnicas de classificação de sentimentos. 56
Figura 11: Padrões de etiquetagem para extração de duas palavras das críticas. 62
Figura 12: Análise de vários fatores sobre o iPhone X (Fonte: Creative Strategies, n=1746, 2018). 69
Figura 13: Índice de Satisfação sobre o tema brexit retirado da aplicação OpClassify implementada e apresentada no Capítulo 4.....70
Figura 14: Metodologia CRISP-DM.....79
Figura 15: Visualização do ficheiro views.py 81
Figura 16: Visualização do ficheiro views.py – funções preprocess, tokenize, e elemento regex. 83
Figura 17: Visualização do ficheiro views.py - função clean_tweet. 84
Figura 18: Visualização do ficheiro views.py - Display dos resultados dos dados analisados. 85
Figura 19: Exemplo da análise de polaridade usando a biblioteca TextBlob. 86

Figura 20: Página inicial da aplicação OpClassify 87
Figura 21: Inserção do tema "brexit" com amostra 10 e loader do sistema em execução. 88
Figura 22: Exemplo de pesquisa do tema "brexit" no OpClassify. 89
Figura 23: Análise dos Tweets Positivos relativamente à pesquisa sobre o tema "brexit". 90
Figura 24: Análise dos Tweets Neutros relativamente à pesquisa sobre o tema "brexit". 91
Figura 25: Análise dos Tweets Negativos relativamente à pesquisa sobre o tema "brexit". 92
Figura 26: Arquitetura implementada para linha cronológica baseada na análise por mês/ano. 93
Figura 27: 1 - Índice de negatividade do dia anterior; 2 - Índice de neutralidade do dia anterior; 3 - Índice de positividade do dia anterior; 4 - Índice de negatividade de 2 dias anteriores; 5 - Índice de neutralidade de 2 dias anteriores; 6 - Índice de positividade de 2 dias anteriores; 7 - Índice de negatividade de 3 dias anteriores; 8 - Índice de neutralidade de 3 dias anteriores; 9 - Índice de positividade de 3 dias anteriores. 95
Figura 28: Módulo de mineração Web padrão (Fonte: https://www.clips.uantwerpen.be/pattern). 97
Figura 29: Análise padrão do tema "brexit". 100
Figura 30: Análise Naive Bayes do tema "brexit". 101
Figura 31: Análise padrão do tema "facebook". 102
Figura 32: Análise Naive Bayes do tema "facebook". 103
Figura 33: Análise padrão do tema "corruption". 104
Figura 34: Análise Naive Bayes do tema "corruption". 105
Figura 35: Análise padrão do tema "ryanair". 106
Figura 36: Análise Naive Bayes do tema "ryanair". 107
Figura 37: Análise padrão do tema "mcdonalds". 109
Figura 38: Análise Naive Bayes do tema "mcdonalds". 110

1 Introdução

"A mineração de opiniões e sentimentos é o estudo computacional das opiniões, atitudes e emoções das pessoas perante uma entidade." (Medhat, Hassan and Korashy, 2014)

1.1 Contextualização

A análise de sentimentos ou mineração de opiniões, pode ser vista como o estudo computacional das opiniões, avaliações, atitudes e emoções das pessoas, em relação a entidades, indivíduos, questões, eventos, tópicos e os seus respetivos atributos. (Liu and Zhang, 2012)

Fundamentalmente, verifica-se que o processo de análise de sentimentos é tecnicamente desafiante, mas também repleto de enorme utilidade prática. Em qualquer tipo de negócio, é recorrente a existência da vontade de saber o que um dado conjunto de clientes (público-alvo) perspetivam, em relação à opinião, visão, e ponto de vista que apresentam relativamente ao uso de um serviço ou de um qualquer produto adquirido.

Além do mais, e cada vez mais, os clientes procuram conhecer e obter o máximo de informação no que concerne às opiniões de outros utilizadores que adquiriram no passado o produto que estes agora pretendem adquirir.

(Liu, 2010) afirmou que a informação textual existente no mundo pode ser categorizada em dois principais tipos: factos e opiniões.

Os factos são expressões objetivas sobre entidades, eventos e as suas propriedades, enquanto que as opiniões são usualmente, e na sua maior parte,

expressões subjetivas que descrevem os sentimentos das pessoas, avaliações ou sentimentos perante entidades, eventos e as suas propriedades.

Uma das razões pelas quais existe uma lacuna no estudo das opiniões é porque antes do aparecimento da *World Wide Web (Web)* existia pouco texto orientado às opiniões.

Nessa altura, quando um indivíduo pretendia tomar uma decisão perguntava, tipicamente, a amigos ou família a sua opinião. No caso de uma organização, esta quando queria conhecer a opinião dos seus clientes sobre os seus produtos promovia a realização de inquéritos, e outro tipo de pesquisas para obter essa informação.

De facto, a *Web* alterou a forma como as pessoas expressam os seus pontos de vista e as suas opiniões. Hoje em dia, qualquer pessoa pode postar críticas dos produtos que comprou a uma dada empresa e expressar os seus pontos de vista sobre quase tudo em fóruns, em grupos de discussão ou em blogs. Esta nova forma de expressão de opiniões traduz um comportamento que permite definir um conjunto de novas métricas sobre a informação, e sobretudo provoca o aparecimento de novas fontes de informação com várias aplicações práticas.

Atualmente se um qualquer indivíduo pretender adquirir um produto, e, antes da aquisição quiser saber mais sobre esse produto, irá, muito provavelmente, consultar as críticas existentes na *Web* sobre esse mesmo produto.

Do ponto de vista de uma empresa, hoje já não é também necessário despender de recursos próprios para conhecer a opinião dos seus clientes sobre os produtos que comercializa, ou para saber as opiniões do mercado sobre os produtos de qualquer empresa sua concorrente.

Com o aumento significativo do *social media,* veja-se os casos da existência de inúmeros *blogs* de *críticas*, fóruns de discussão, e redes sociais (*Facebook, Instagram, Twitter, etc.*), é de notar que os indivíduos ou as organizações beneficiam do efeito e da influência que as opiniões públicas expressas em todo

o tipo de meios acima citados têm sobre o estruturar das suas decisões destes mesmos indivíduos ou nestas mesmas organizações.

Embora a quantidade de dados gerada por toda essa atividade opinativa exercida no contexto de mercado seja muito grande e enormemente rica, facilmente se constata que a pesquisa, organização, monitorização e interpretação desses dados, para ter efeitos benéficos junto dos objetivos empresariais, são operações muito complexas e trabalhosas. Isto porque, para além da tarefa de encontrar estes mesmos dados, sobrepõe-se a tarefa de tratar e, mais especificamente, interpretar os dados obtidos.

Além disso, acresce também a complexidade da análise de tais dados. Basta observar que, para cada sítio *Web*, pode encontrar-se um enorme volume de texto com características e perspetivas, na sua maioria, bastante específicas. Quer isto dizer, que, na maior parte dos casos, as opiniões expressas nestes sítios são bastante intricadas e difíceis de interpretar, com vista à sua categorização e determinação da sua utilidade.

Mesmo ao nível do "olho humano", aquando da leitura e pesquisa nos sítios referidos contendo inúmeras *críticas*, algumas destas escapam e a sumariação da sua informação revela-se de difícil execução. Consequentemente, o conhecimento retirado da pesquisa e análise das *críticas* pode não ser o melhor possível ou mesmo não ser útil. Além disso, com um volume tão significativo de informação a retirada de conhecimento útil dessa mesma torna-se uma tarefa deveras complexa.

Tendo isto em conta, a mineração de opiniões e de sentimentos acrescida de um sistema de sumarização é uma necessidade, uma vez que daí advêm enormes benefícios a diferentes níveis, tanto para consumidores como para organizações em termos gerais.

1.2 Motivação e Objetivos

A análise de sentimentos ou a mineração de opiniões referem-se ao estudo e consequente tentativa de interpretação computacional das opiniões, atitudes e emoções das pessoas (Medhat, Hassan and Korashy, 2014).

Tais aspetos permitem revelar entidades, definindo perfis, que possibilitam a execução de processos de *profiling* sobre as próprias entidades ou sobre os indivíduos que constituem.

Este tipo de estudo computacional sobre perfis permite traçar planeamentos de negócio mais efetivos e melhor fundamentados, caso sejam realizados sobre indivíduos que constituam o público-alvo desse negócio. Tomemos o exemplo de várias operadoras de comunicações nacionais que efetuam, após uma chamada de suporte, um inquérito ao cliente sobre o grau de satisfação que possui relativamente ao serviço/atendimento prestado.

Tudo isto é informação que permite a criação de bases sólidas para melhorar os serviços prestados aos clientes. Como tal têm como objetivo melhorar a performance da empresa no mercado, conservando-a num bom nível de satisfação dos seus clientes, cativando possivelmente novos clientes.

Também existem exemplos de sítios *Web* de críticas (Pang and Lee, 2008) sobre hotéis e restaurantes, que permitem fazer inúmeras críticas benéficas, quer para quem quer conhecer a experiência e tomar uma decisão sobre um hotel ou uma viagem que querem comprar, quer para as entidades que prestam serviços ou vendem os produtos que estejam presentes no mercado (Dellarocas, Zhang and Awad, 2007) e que são usualmente passíveis de serem avaliados.

No mercado financeiro (Devitt and Ahmad, 2007) a identificação da polaridade dos sentimentos existentes nas opiniões partilhadas em plataformas sociais ou em sítios *Web* de críticas já provou, de igual modo, ser de grande influência na venda ou compra de ações de uma dada companhia, consoante a

melhor cotação sentimental obtida nas várias críticas existentes e analisadas pelos interessados no mercado.

Todos estes sítios *Web* procuram proporcionar a melhor informação e conhecimento ao cliente, de forma a que este possa tomar a melhor decisão.

Desta forma, um negócio pode suportar grande parte das suas melhorias baseando-se na interação com o cliente, tentando assim aprofundar a relação empresa-cliente para saber quais os aspetos de negócio que pode melhorar ou quais os aspetos que o cliente gostaria que fossem resolvidos, por exemplo.

Grandes benefícios advêm, não só, da análise de sentimentos e mineração de opiniões, como, em adição a estes mesmos processos, a utilização dos dados obtidos no estabelecimento de índices de satisfação. Este tipo de índices permite observar os cenários envolventes no que diz respeito à satisfação de consumidores relativamente a dado produto ou serviço proporcionado por uma empresa ou instituição. (McGlohon, Glance and Reiter, 2010)

Assim, podemos concluir que a mineração (e análise) de opiniões e de sentimentos pode gerar um grande movimento das massas na nova indústria 4.0, uma vez que as ações de compra e venda, por exemplo, são facilmente influenciadas por opiniões assumidas através dos diferentes meios tecnológicos e pela informação existente, de modo que bom processo de mineração de opiniões e de sentimentos pode antecipar uma posição no mercado mais favorável para um dado negócio ou evento.

Neste trabalho de dissertação abordámos a temática da mineração de opiniões e análise de sentimentos, como base para o estabelecimento de índices de satisfação, motivados pelo relevo dos resultados que alguns dos trabalhos publicados no domínio alcançaram. Veja-se por exemplo os casos de (Zhan, Loh and Liu, 2009), (Ortigosa, Martín and Carro, 2014), (Anjaria and Guddeti, 2014) e (Breault, Goodall and Fos, 2002).

(Zhan, Loh and Liu, 2009) estudaram o impacto das críticas sobre produtos nas preocupações dos consumidores-alvo de um produto (possíveis consumidores futuros de um produto) e como estas revelavam fielmente a

experiência dos consumidores com o produto em causa. Esta informação é essencial, tendo em conta a sua benéfica utilização para propósitos de arquitetura conceptual, personalização, melhoria de serviços, recomendação de produtos entre consumidores, melhor compreensão dos desejos dos consumidores, e finalmente para melhoria da atração de novos consumidores para o negócio. Vários estudos que especificam a derivação de grandes quantidades de dados sob a forma de críticas de consumidores focam-se, principalmente em dados numéricos e categóricos.

Contudo, os dados textuais contêm um valor bastante alto no que toca à posse de informação em si propriamente dita, e, apesar da complexidade do seu processamento e consequente análise, são uma classe de dados que permite, de forma mais aprofundada, conhecer de perto as opiniões e sentimentos dos consumidores. (Zhan, Loh and Liu, 2009) afirmam a existência de vários métodos de mineração de opiniões no que toca ao processamento de críticas de produtos e concluem que a maioria se concentra apenas na contabilização de comentários positivos e negativos por partes dos autores das críticas, o que, afirmam, não é o suficiente para dar cobertura a todos os tópicos, preocupações e elementos cruciais e transversais a todo o conjunto de dados que formam o conjunto de críticas sob análise.

Ao invés disso mesmo, os autores (Zhan, Loh and Liu, 2009) propõem uma abordagem fundamentada na sumariação automática baseada na análise da estrutura interna de tópicos dos artigos opinativos em análise de modo a reunir todos os pontos abordados pelas perspetivas opinativas dos consumidores.

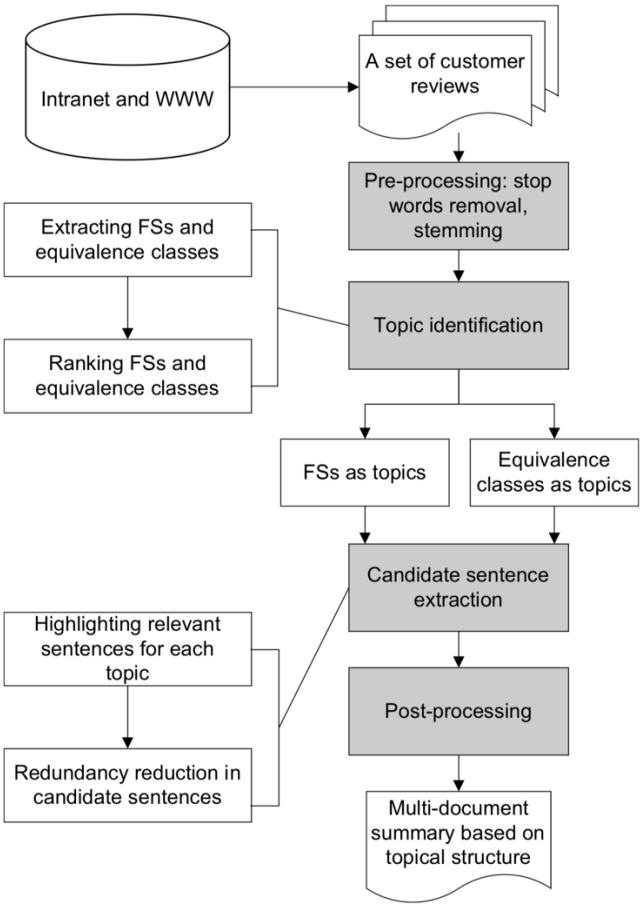


Figura 1: Processo de Sumariação baseado na estrutura dos tópicos.

(Zhan, Loh and Liu, 2009) diferenciaram-se dos métodos existentes para a construção das perspetivas de sumariação centradas na hierarquização e agrupamento das frases em análise, sendo que o seu método é descobrir e extrair os tópicos salientes de todo um conjunto de críticas e após tal proceder à hierarquização dos tópicos encontrados. O sumário final é assim retirado e gerado tendo por base a hierarquia dos tópicos, como se pode visualizar na Figura 2 tendo por base o estudo do produto "*Nokia 6610*".

O estudo demonstrou que a técnica utilizada obteve sucesso em termos do que os consumidores consideravam tópicos relevantes em relação a determinado produto e em relação à capacidade desta mesma técnica de descobrir os tópicos basilares de cada produto que foi analisado.

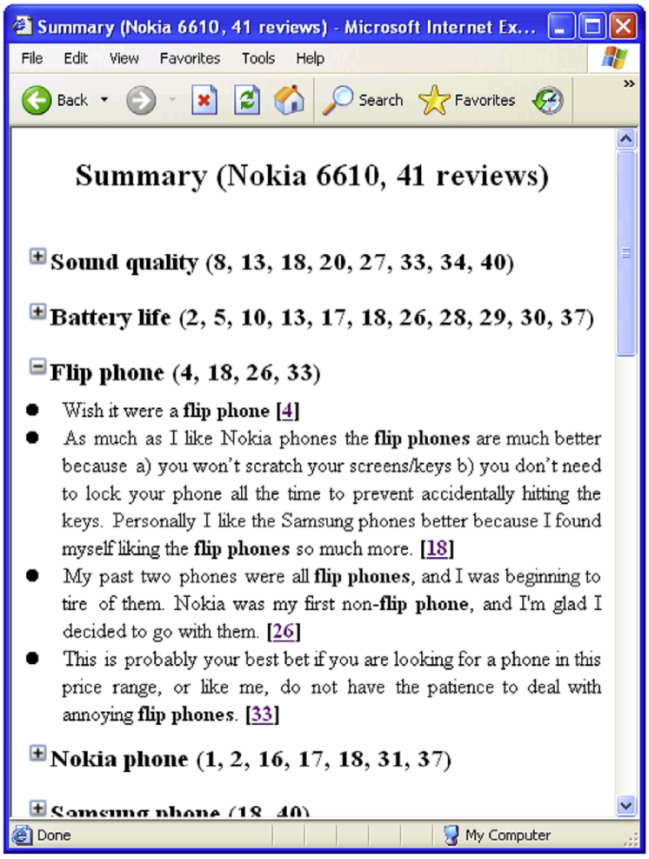


Figura 2: Resultado da sumariação para o conjunto de críticas sobre o Nokia 6610.

(Ortigosa, Martín and Carro, 2014) apresentaram um artigo onde desenvolvem um novo método sobre a análise de sentimentos na rede social *Facebook*, e que suporta a extração de informação tendo em conta a polaridade de sentimentos das mensagens escritas pelos utilizadores, e que implementa um modelo baseado na polaridade de sentimentos revelada pelos utilizadores e que permite detetar alterações emocionais significativas. Este método foi implementado com recurso ao *SentBuk*, sendo que esta aplicação retorna as mensagens escritas pelos utilizadores na rede social e classifica-as de acordo com a sua respetiva polaridade, mostrando os resultados para cada utilizador através de uma interface interativa, como se pode verificar na Figura 3.



Figura 3: Aplicação SentBuk

Outra das principais características em relevo é a deteção de alterações emocionais significativas, pesquisa de emoções dos amigos conectados ao utilizador, classificação de utilizadores de acordo com as suas mensagens e respetivas estatísticas, entre outras. O método de classificação implementado na aplicação SentBuk baseia-se numa abordagem híbrida: combina técnicas de Machine Learning com técnicas baseadas na interpretação do léxico presentes nas mensagens que são analisadas pela aplicação.

Os resultados obtidos através desta abordagem mostram que é uma técnica satisfatória em relação à execução de análise de sentimentos no *Facebook* com alta percentagem de sucesso (83.27%).

No artigo é dado relevo a uma importante aplicação desta mesma técnica, mais propriamente no contexto do *e-learning* e criando um sistema de aprendizagem personalizado, analisando que tipo de problemas devem ser colocados ao utilizador num contexto de aprendizagem no qual o sistema tem em conta as emoções momentâneas do utilizador, isto é, a cada instante da aprendizagem que se encontra a ocorrer. Mais do que isso, este sistema pode também ser utilizado como feedback para professores no contexto de

aprendizagem *online*. Toda esta aplicabilidade do sistema é apresentado no trabalho desenvolvido por (Ortigosa, Martín and Carro, 2014).

(Anjaria and Guddeti, 2014) recorrem ao *Twitter* como fonte de dados, afirmando desde logo que esta plataforma é um ponto de atração para vários investigadores que estudam a influência dos comentários feitos pelos utilizadores da plataforma sobre eventos eleitorais, vendas dos produtos de diversas marcas, lucros na indústria do cinema, lucros na bolsa de valores, popularidade de atores, políticos, entre outros. O Twitter representar uma rápida e eficaz maneira de monitorizar os sentimentos públicos, os sentimentos que abundam numa sociedade num determinado espaço de tempo. No trabalho desenvolvido por (Anjaria and Guddeti, 2014) é apresentada uma abordagem que explora o fator de influência dos utilizadores do *Twitter* e as respetivas mensagens por ele produzidas, com o objetivo de compreender o peso que tais fatores representam para o resultado de eleições. É proposto também um modelo híbrido para extração de opiniões utilizando características diretas e indiretas patentes nos dados presentes no Twitter, baseado em SVM, Naive Bayes, Entropia Máxima e classificadores supervisionados baseados em redes neuronais artificiais. Foi realizada uma combinação entre a análise de componentes principais (PCA) e SVM na tentativa de executar a redução de dimensionalidade.

Tal método foi aplicado em dois cenários sociais diferentes, o primeiro foram as eleições presidenciais dos Estados Unidos em 2012 e nas eleições para a assembleia de Karnataka[1] em 2013. Os resultados experimentais demonstram que o modelo SVM obteve maior taxa de sucesso do que qualquer outro modelo com uma precisão máxima de sucesso a atingir os 88% no caso das eleições presidenciais dos Estados Unidos em Novembro de 2012 e com uma precisão máxima de sucesso de 58% no caso das eleições para a assembleia do estado de Karnataka em Maio de 2013.

[1] Karnataka é um dos estados da Índia, localizado no sul do país.

(Breault, Goodall and Fos, 2002) focaram-se num problema de saúde frequente em todo o mundo, a diabetes. Como fonte de dados para o desenvolvimento do seu trabalho utilizaram várias bases de dados e registos relativas a pessoas que sofrem de diabetes com informação recolhida sistematicamente. Analisaram todos estes dados presentes num Data Warehouse de um grande sistema de cuidados de saúde intensivos em Nova Orleães nos Estados Unidos que tem dados registados de cerca de 30,383 diabéticos. A primeira dificuldade relatada no artigo é a translação de todos os dados para um ficheiro capaz de albergar condições para a execução de mineração de dados. Tomaram como referência duas variáveis especificamente relacionadas com a diabetes, um índice de comorbidade e o valor de HgbA1c[2]. Foram utilizadas árvores de decisão com a variável binária relativa ao valor 9.5 de HgbA1c e 10 fatores de previsão: idade, sexo, visitas ao departamento de emergência, visitas ao médico, índice de comorbidade, dislipidemia[3], hipertensão, doenças cardiovasculares, retinopatia[4], e doença renal em estágio final. Surpreendentemente, a variável mais importante associada com um mau controlo glicémico é a idade jovem, e não o índice de comorbidade ou se os pacientes têm outras doenças relacionadas. Os piores valores de HgbA1c são 3.2 vezes mais fáceis de encontrar em doentes com menos do que 6.5 anos de idade.

Existem inúmeras aplicações para a análise de opiniões e sentimentos, desde aplicações que prevêem a realização de um empréstimo bancário com base na idade, situação profissional entre outros fatores de certa pessoa, até a aplicações que calculem com quantificada certeza a predisposição de certo indivíduo em relação a certa patologia. É, indubitavelmente, um campo de estudo

[2] Forma de hemoglobina presente naturalmente nos eritrócitos humanos que é útil na identificação de altos níveis de glicemia durante períodos prolongados. Este tipo de hemoglobina forma-se a partir de reações não enzimáticas entre a hemoglobina e a glucose.
[3] Dislipidemia ocorre quando se verificam anomalias quantitativas ou qualitativas dos lípidos (gorduras) no sangue.
[4] Retinopatia diabética é uma lesão da retina (estrutura transparente e sensível à luz localizada na parte posterior do olho) que tem como causa o diabetes.

que potencia sem sombra de dúvida, o progresso em muitos outros campos de estudo existentes, seja progredindo na investigação, seja progredindo a nível de procedimentos, organização de estruturas, ou a nível de conhecimentos do campo em causa.

Com este trabalho pretende demonstrar-se que a mineração de opiniões e sentimentos providencia uma perspetiva bastante proveitosa sobre qualquer setor de mercado em específico, de modo a fornecer excelentes fundamentos para a melhor tomada de decisões possível.

Seguidamente enumera-se um conjunto de objetivos que se pretendeu alcançar com o desenvolvimento deste trabalho:

- Aumento significativo da excelência da fundamentação das decisões tomadas no âmbito empresarial como forma de construir o futuro de uma empresa;
- Incremento do fornecimento do número de perspetivas que certo tema possui no mercado, essencialmente a nível social, seja para a empresa que atua no mercado ao efetuar um estudo de mercado entre outros, seja para o consumidor que deseja efetuar determinado estudo de mercado e/ou aquisição de certo produto/serviço entre outros;
- Vista cronológica do efeito causado por certo tema/produto/serviço no mercado.
- Perspetiva cronológica fornecida sob três parâmetros: negatividade, neutralidade, positividade, ou por índice *opclassification*[5].
- Quantidade da amostra recolhida variável, sendo que implementa um modo dinâmico de obtenção de resultados. É permitida a afinação da granularidade pretendida, sendo que uma amostra maior resulta, na maior parte dos casos, numa maior fidelidade dos resultados obtidos e analisados.

[5] Índice *opclassification* explanado no subcapítulo 5.8.

- Para cada um dos fatores analisados nomeadamente positividade, neutralidade e negatividade, foi apresentada uma análise pormenorizada na medida em que é disponibilizada uma *wordcloud* para um dos fatores, e, além disso, é disponibilidade para cada frase pertencente a cada um dos fatores os respetivos resultados referentes à polaridade e à subjetividade.
- Possibilidade de retirada de ilações fundamentadas mediante e baseadas nos resultados apresentados, disponibilizando deste modo uma referência para toda uma panóplia de cenários do mercado em causa, permitindo uma melhor análise do mercado antes da tomada de qualquer decisão por parte da empresa/consumidor.

1.3 Estrutura do documento

Para além do presente capítulo, esta dissertação está organizada em mais cinco capítulos, nomeadamente:

- Capítulo 2 – Mineração de Opiniões e Sentimentos – neste capítulo aborda-se o domínio referente à Mineração de Opiniões e Sentimentos, expondo uma primeira abordagem ao conceito inerente a este tema. É explanado todo o sistema existente para suporte desta mesma mineração, e todos os benefícios que dele mesmo advêm. Ainda neste capítulo é constituída e exposta uma linha cronológica dos progressos realizados neste campo e todos os desenvolvimentos que

foram feitos ao longo do tempo tendo em conta o avanço deste mesmo sistema de Mineração de Opiniões e Sentimentos. A exposição desta mesma linha cronológica é dividida em 5 fases organizadas em subcapítulos, respetivamente, sendo que, em cada subcapítulo, são descritos os avanços realizados nessa mesma fase temporal sobre este mesmo campo.

- Capítulo 3 – Mineração de Opiniões e Sentimentos para Estabelecimento de Índices de Satisfação – neste capítulo é realizada a formulação do problema em si propriamente dita. É inicialmente apresentada uma definição de mineração de opiniões e sentimentos e como tal pode servir para estruturar, fundamental e implementar Índices de Satisfação na tentativa de disponibilizar intuitivamente os resultados desta mesma mineração sobre os elementos específicos nomeadamente opiniões e sentimentos.

- Capítulo 4 – Análise de Opiniões e Sentimentos – neste capítulo a Análise de Opiniões e Sentimentos é abordada sendo primeiramente apresentada uma definição do que é este tipo de análise e a que se refere especificamente. É feita toda uma contextualização deste tema tendo em conta a sua fundamental importância em todo um sistema de análise relativo a este campo. No primeiro subcapítulo são distinguidos diferentes tipos de análise, nomeadamente três: a nível documental, a nível frásico, e a nível de aspetos. O léxico utilizado relativamente ao campo sentimental do ser humano é abordado no segundo subcapítulo. No terceiro subcapítulo analisamos ao pormenor a matéria-prima de todo um sistema a construir neste campo em específico: as opiniões e os sentimentos. Seguidamente, a polaridade dos dados é descrita e explicada no quarto subcapítulo onde se revela a preponderância de tal fator num sistema de análise de opiniões e

sentimentos. No quinto subcapítulo designado Classificação de Sentimentos distinguimos dois subcapítulos dentro deste mesmo, nomeadamente a Classificação de Sentimentos baseada em Aprendizagem Supervisionada e a Classificação de Sentimentos baseada em Aprendizagem Não Supervisionada. O último e sexto subcapítulo inserido no quarto capítulo denomina-se de Índices de Satisfação onde é abordada a fundamental importância e o papel que este tipo de índices possuem em diferentes tipos de cenários.

- Capítulo 5 – Caso de estudo – neste mesmo capítulo é realizada a explanação de todo o processo de arquitetura, planeamento, recolha de requisitos, idealização do caso de estudo, listagem dos objetivos a atingir com o sistema em causa, implementação, especificação de cada processo inerente à implementação deste mesmo sistema, disponibilização do produto final, e descrição dos resultados obtidos em diferentes cenários e com diferentes modelos utilizados no sistema.

Por fim, todas as referências utilizadas neste documento são descritas e enumeradas.

2 Mineração de Opiniões e Sentimentos

Na aldeia global em que vivemos existe um enorme património de textos contendo opiniões, sentimentos, sobre um determinado acontecimento, bem ou serviço. Todas as redes que estão criadas e são sustentadas permitem um enorme crescimento significativo no que concerne à quantidade de dados existentes sobre textos, uma vez que permitem que quase todo o tipo de opiniões seja expresso.

Cada utilizador de cada rede expressa a sua opinião contribuindo para uma massificação do "oceano" de dados existente, e que pode, sem dúvida alguma, constituir as bases de um estudo estruturado em vários eixos sobre o seu conteúdo. Várias perspetivas de análise podem ser desenvolvidas sobre o leque de cenários em que um vultoso volume de dados gerados pela comunidade em causa é útil quando interpretado como um índice de satisfação, podendo mesmo atuar como alavanca para uma melhoria de serviços, tendo em conta o incremento de performance junto das atividades exercidas para o alvo do mercado, o cliente. (Lee, Jeong and Lee, 2008)

As redes sociais, nomeadamente o *Twitter* (Liu, 2012) e o *Facebook*, constituem marcas incontornáveis da independência do pensamento. Todos os dias, são gerados em forma de tweet (no *Twitter*) ou estado (no *Facebook*), grandes volumes de opiniões, reveladoras de sentimentos, bons ou maus, por parte de quem efetua a publicação. O peso influenciador que as várias plataformas de ação e partilha social podem ter na opinião das pessoas, neste caso mais específico, o *Youtube*, obriga as próprias marcas, muitas das vezes, a apostar mais na publicidade e na abertura de portas através de vídeos para com os utilizadores da *Web*, expondo de modo mais pormenorizado os seus produtos e a forma como estes são feitos, enlevando todo o processo de manufatura associada à produção dos mesmos. Todos estes fatores associados à propagação

que uma dada *review* pode ter, obrigam a um reposicionar das ações que as marcas, empresas e indústrias protagonizam, em relação a como apresentam um produto, e principalmente a uma maior restrição e disciplina das normas de atendimento a um cliente, por exemplo. Cada cliente desempenha um papel fundamental na ação de uma empresa, junto do seu público-alvo, uma vez que a partilha de experiências sobre algo comercializado, hoje em dia, é muito maior. Quer isto dizer que, o que para uma marca pode parecer e ser apenas um "simples" cliente ou "apenas mais um cliente" pode rapidamente tornar-se no cliente que produz um vídeo em que, no caso de incorreto atendimento ou incorreta prestação de serviços por parte da empresa em questão, expõe o seu relato negativo com a empresa numa plataforma social, levando a que muitas mais pessoas partilhem, partilhem a sua opinião e tenham conhecimento do sucedido, afetando assim, naturalmente, a imagem da empresa em questão.

Uma empresa posicionada no mercado da venda de produtos pode ser avaliada de acordo com vários eixos de análise, relativamente aos serviços que presta no mercado em que está inserida. Hoje, os clientes já fazem, de forma bastante regular, a avaliação de produtos e de serviços obtidos, fornecidos e prestados por uma empresa. Várias das atividades promovidas pelos clientes, como a publicação de *reviews* sobre os produtos e sobre serviços que adquirem através de certa empresa ou a discussão em fóruns ou *blogs/videoblogs* sobre os produtos que adquirem, entre outras, geram discussão e consequentemente a criação, a edição e a mudança de opinião em relação a um produto (ou produtos) de uma empresa. Assim, adianta notar que a imagem que uma empresa transmite para o mercado em que atua, não é "tudo", mas, mesmo assim, representa quase "tudo" no que diz respeito ao posicionamento da empresa nesse mesmo mercado.

Uma empresa bem cotada é, significativamente e idealmente, uma empresa que apresenta críticas positivas dos seus produtos nos mais variados segmentos de redes nas quais existe a hipótese de expressar, de forma individual ou coletiva, a avaliação de um produto ou de um serviço prestado por uma empresa. Assim,

ao não conseguir obter uma totalidade positiva de comentários e de críticas sobre um dos seus produtos, resta à empresa obter o máximo de enquadramento positivo no mercado. Quer isto dizer que, uma empresa deve empregar todo o esforço em servir bem os seus clientes, para assim provocar a criação de uma inflação positiva da sua atividade empresarial no seio do mercado em que desenvolver as suas atividades. Dito isto, uma empresa tem grandes vantagens quando possuiu e utiliza ferramentas de aprendizagem automática para extrair conhecimento útil através dos dados extraídos das opiniões e sentimentos revelados pelos seus clientes.

No nosso quotidiano assiste-se, diversas vezes, a tomadas de ação por parte de empresas para com alguns sujeitos individuais, dado que estes manifestaram em plataformas sociais o seu descontentamento para com o serviço que obtiveram por parte dessa mesma empresa, sendo que o seu manifesto atingiu um grande número de "*shares*" levando a uma danificação da imagem da empresa. Por vezes, com o elevado número de "shares" atinge-se uma exposição mediática ainda maior de certos casos quando a review atinge os próprio *media*. Quando esses momentos são atingidos notam-se atitudes responsivas por parte das empresas, como uma reparação gratuita do ecrã de um *smartphone* pelo qual pediam um valor excessivo, ou até a correta reparação do *smartphone*, uma vez que o repararam, mas não com peças originais da sua marca.

Também já foi provado que 81% dos utilizadores da *Internet* (Pang and Lee, 2008), antes de comprar um produto, procuram opiniões sobre esse mesmo produto, de forma a poderem formar uma opinião mais sustentada que os conduza a uma decisão (Tiwari, Turner and Majeed, 2008), ou seja, estas pessoas são influenciados pelas opiniões, sentimentos, emoções, experiências de outros utilizadores do produto, que partilharam a sua experiência na forma de opinião, sentimento ou emoção em algum sítio da rede. Estes comportamentos fornecem dados significativos, que não são despiciendos e merecem ser devidamente analisados para que através deles, se possa obter informação que

possa ajudar num processo qualquer de tomada de decisão (Bhuiyan, Xu and Jøsang, 2009). Mesmo com o aparecimento da *Web 2.0* (Ravi and Ravi, 2015), mais concretamente, com a massificação de opiniões, assistimos aos constantes aparecimentos de *blogs*, *microblogs* (Liu, 2012) e canais de *Youtube*, nos quais muitas das vezes um indivíduo é seguido por milhares (ou mesmo milhões) de outros indivíduos que o tomam como referência para formular opiniões e tomar decisões. Esta "indústria" de análise de dados permite, em vários pontos, retirar ilações sobre as perspetivas do público-alvo de um dado negócio ou de um dado acontecimento, de forma a corresponder às suas expectativas, melhorando desse modo a satisfação geral desse mesmo público-alvo. Ou seja, mais do que melhorar, é possível antecipar um cenário de êxito, tendo em conta a satisfação que um dado público-alvo procura (Zhang and Zhu, 2013). Estas técnicas de mineração de dados enfrentam assim *tweets* e estados diariamente publicados, tentando por essa via, retirar, na medida do possível, uma avaliação positiva, negativa ou neutra a partir dos dados gerados pelos utilizadores de redes sociais. Tendo em vista o desafio de "ler" e "prever" os comportamentos desses mesmos utilizadores, é necessário que exista uma técnica bem fundamentada para a mineração de opiniões e sentimentos (Somprasertsri, 2010). Tal facto, justifica-se com o objetivo de estabelecer determinadas correntes de pensamento que possam ser interpretadas através desses dados, e, acima de tudo, fazer a geração de eventuais índices de satisfação, não só para o seu público-alvo, mas também para as estruturas de negócios que estão sob constante avaliação do mercado.

2001 foi um marco no estudo das opiniões e sentimentos de seres humanos e na obtenção de conhecimento útil (Pang and Lee, 2008). Se seguirmos uma linha cronológica (Liu, 2012), podemos ver que um modelo de sentimentos foi proposto com o objetivo de prever a performance de vendas (Liu, 2012). Como prova, podemos ver que em:

- (McGlohon, Glance and Reiter, 2010), as *reviews* foram usadas para organizar um ranking de produtos e de comerciantes.

- (Hong and Skienay, 2010), as relações entre a linha de apostas na NFL e as opiniões públicas em blogs e no *Twitter* foram estudadas.
- (O'Connor *et al.*, 2010), os sentimentos revelados no *Twitter* foram correlacionados com pesquisas de opiniões públicas.
- (Tumasjan *et al.*, 2010), os sentimentos revelados através de tweets foram também cotejados com as previsões em relação a resultados eleitorais específicos.

A ferramenta usada para estas análises foi o software de análise de texto *LIWC (Linguistic Inquiry and Word Count)*, tendo suportado uma análise de conteúdos sobre mais de 100 000 mensagens, contendo referências em relação a partidos políticos, ou, mais especificamente, a figuras políticas. Os resultados obtidos mostram que o *Twitter*, é, na verdade, usado extensivamente para refletir o resultado das eleições e que, as publicações feitas nessa mesma rede, podem, de facto, espelhar inclinações sobre certas previsões do resultado de eleições políticas específicas. Para além disso, foi notado que o mero número de mensagens referindo uma parte política foi, na verdade, um espelho do resultado das eleições. Mais do que isso, menções conjuntas de duas partes através de *tweets* estão em linha e concordantes com as alianças e coligações políticas reais. Além disso, uma análise dos *tweets* que contêm sentimentos políticos demonstraram uma correspondência restrita e concisa, quando comparamos essa mesma linha ideológica com a linha definida pelo cenário central em causa, um cenário que é, de certo modo, delineado pelas partes e figuras políticas através das suas posições políticas. Por isso, tal facto indica que o conteúdo dos tweets reflete, plausivelmente, o cenário político existente no modo *offline*. Os autores desses trabalhos adiantam, ainda, que o objetivo da exposição é, de facto, discutir a possível utilização das mensagens de *microblogging* como um indicador válido de sentimentos políticos por parte dos cidadãos, e, acima de tudo, derivar sugestões para futura investigação.

Além destes trabalhos, em 2012, (Soler, Cuartero and Roblizo, 2012) usaram as eleições espanholas para investigar o uso do Twitter com o mesmo propósito: a explicitação de decisões pessoais de cada indivíduo em relação a uma eleição, sobre uma dada perspetiva política. Desse trabalho retiraram-se muitas conclusões, entre elas a de que, as conversações mantidas na rede do Twitter sugerem antecipações em relação à decisão de voto de cada indivíduo, e, de alguma forma, antecipam de igual modo os resultados das eleições. Com tal objetivo definido, o grupo de pesquisadores desenvolveu uma ferramenta chamada *Taratweet* para assim definir experiências e capturar as conversas definidas. As experiências tiveram em conta 3 eleições ocorridas entre 2011 e 2012. Os resultados dessas eleições mostram que o Twitter é, de facto, utilizado para discussões políticas e que as referências às diferentes partes constituintes do cenário político correspondem, significativamente, aos votos dos eleitores. Este é um muito possível indicador de que o Twitter pode ser objeto de estudo por parte de futuros investigadores como uma ferramenta de acrescida expressão social em várias dimensões, que permita prever os resultados de futuras eleições. Assim, é notório que a mineração de opiniões e de sentimentos é, na verdade, uma fonte das verdadeiras vontades, das verdadeiras expressões de todo o utilizador de redes de expressão social. Deste modo, retirar ilações após um correto tratamento e uma consequente interpretação dos dados apresenta várias capacidades de extração de conhecimento útil com utilidade para várias áreas aplicacionais.

2.1 Linha cronológica

Para uma melhor contextualização de um possível estado da arte no domínio da mineração de opiniões e de sentimentos será proveitoso observar os seus diferentes pontos de vista da mesma, bem como notar o trabalho desenvolvido no seu seio (Binali, Potdar and Wu, 2009). Prova disso mesmo, e constituindo outra perspetiva do estado da arte, muito do trabalho realizado até hoje nesse campo pode ser observado na exposição realizada por (Anbananthen, Kalaiarasi, & Elyasir, 2013)(Sonai Muthu Anbananthen, Kalaiarasi and Mohamed H. Elyasir, 2013). Estes autores encetam por afirmar que as opiniões são informação que carrega uma quantidade significativa de valor interpretativo, uma vez que é de fundamental importância o poder de saber qual o objeto de pensamento ou a movimentação social. De facto, saber o que as pessoas falam entre si, e, ainda mais importante, conhecer os sentimentos desses mesmos indivíduos constitui a base para a construção de poderosas ferramentas para ajudar a justificação e a estruturação de melhores decisões possíveis a tomar. Além disso, esses autores também afirmam que a mineração de opiniões e de análise de sentimentos surgiram como tecnologias desenvolvidas com o intuito de providenciar os melhores métodos para analisar opiniões e sentimentos presentes em textos. Tais tecnologias focam-se não só em retornar conhecimento relevante a partir de documentos, de textos ou de frases, mas também em extrair conhecimento útil a partir das opiniões e sentimentos contidos nesses textos. No artigo referido é estabelecido também o objetivo de estruturar uma linha cronológica da evolução do estudo na área em investigação. Tendo tal em vista, os autores realizaram uma *review* da literatura de artigos publicados desde 1970 até ao presente. Tal análise permitiu estabelecer uma linha de progresso no campo em questão, que foi dividida em 5 fases, nomeadamente:

1) A Interpretação de Texto (1979-1995).
2) Anotação de Opiniões de Baixo-Nível (1997-2003).
3) Diferença entre Subjetividade e Objetividade (2003-2005).
4) Aplicações Web de Mineração de Dados (2006-2007).
5) Recursos Léxicos (2008-2011).

Vejamos, então, de forma mais detalhada, o que cada uma destas fases envolveu.

2.1.1 A Interpretação de Texto (1979-1995)

Todas as pessoas possuem diferentes maneiras de pensar e de agir, cada uma delas baseadas em múltiplos fatores e elementos que incluem, por exemplo, motivações, objetivos, circunstâncias, tempo ou lugar. Estes factos fazem dos seres humanos elementos subjetivos, e, por isso, inconsistentes em alguns aspetos, em particular no que respeita ao domínio da interpretação dos textos que produzem. Em 1979 (Carbonell, J. G., 1979) propôs um modelo de processos centrado em aspetos políticos. A escolha recaiu sobre o domínio da política uma vez que neste domínio existe grande exposição de subjetividade presente no discurso individual e coletivo. Neste modelo a subjetividade das pessoas foi categorizada em duas opiniões: liberal e conservativa. Cada uma destas ideologias ofereceu diferentes interpretações do evento de *inputs* que, obviamente, produziu várias possíveis interpretações.

Esta primeira fase inicia-se aquando do momento em que surge a necessidade de encontrar uma forma de interpretar o texto de forma automática, ou seja, o momento em que se manifesta o objetivo de realizar

um sistema que seja capaz de interpretar um texto tão bem quanto o possível, isto é, tão bem quanto um indivíduo pleno de capacidades no que toca a métricas de interpretação de texto em linguagem natural e a índices de desempenho analítico de texto nestas mesmas condições consegue fazer.

A primeira fase testemunhou o aparecimento da teoria de *Banfield*, *"Discurso e Pensamento Representado"*, teoria esta extensivamente usada nos inícios da mineração de opiniões. Esta teoria propôs a utilização de frases objetivas e subjetivas como indicadores, para pesquisar e inquirir o texto com simples questões, e tomando recurso do elemento psicológico como um importante fator do processamento de linguagem natural. Esta teoria foi usada como modelo linguístico para medir a aceitabilidade ou inaceitabilidade gramática através de intuições e porções de linguagem extraída do narrador nativo. Por sua vez, (Wilks and Bien, 1983) apresentaram um modelo para interpretar a linguagem natural e lidar com a subjetividade do texto. Mais do que isso, assumiram pela primeira vez a possibilidade das Ciências da Computação compreenderem o conceito metafórico e as relações interconectadas entre as frases.

Depois (Wiebe & Rapaport, 1984) propuseram um modelo para a interpretação de perspetivas baseado na teoria de *Banfield*, na qual categorizavam a frase em objetiva ou subjetiva. Uma frase subjetiva representa uma opinião dinâmica ou um comentário que possui a habilidade de mutar de tempos a tempos, enquanto uma frase objetiva retrata factos tais como "A bola é redonda". A interpretação de texto explanada neste modelo tem em conta os pensamentos e percepções do sujeito, no qual reconhece objetividade ou subjetividade na narrativa enquanto terceira pessoa.

2.1.2 Anotação de Opiniões de Baixo-Nível (1997-2003)

As ferramentas de anotação e marcação são extensivamente investigadas nesta fase com o intuito de substituir a tradicional interpretação de texto e minimizar a interação humana para inserção e análise. A anotação de texto introduziu a ideia de marcação e de etiquetagem do texto para futura análise, o que motivou a utilização de classificadores na mineração de dados. Em 1977, (Spertus, 1997) discutiu a ideia inovadora de Chamas (mensagens abusivas) partilhadas online e uma forma para reconhecer a sua existência no interior de um dado contexto online. O nome deste sistema de anotações é *Smokey*. Este sistema desenvolver o seu trabalho em 5 fases principais:

1) As mensagens são convertidas num formato comum com uma frase por linha e delimitadores entre mensagens;
2) O texto passa por um *parser* para anotações;
3) O *output* do *parser* é convertido;
4) As regras de anotações pré-definidas são utilizadas para cada mensagem produzida;
5) A avaliação é feita usando geradores de árvores de decisão.

Além deste trabalho, foram efetuadas outras pesquisas (Howe and Cardie, 1997) para investigar o uso de métodos empíricos em aplicações construídas usando técnicas de linguagem natural. Aqui, o foco centra-se na tarefa de extração e de anotação de informação, nas quais o *input* é um texto não restrito, sendo este mesmo texto é então classificado segundo um tópico predefinido ou domínio de interesse. Posteriormente a informação é sumariada de forma estruturada. A segunda fase do friso cronológico da mineração de opiniões e análise de sentimentos ilustra esta mesma fase como

sendo a etapa de desenvolvimento da área de anotação de texto, que é definida como a área da etiquetagem e de marcação de palavras enquanto a leitura de qualquer tipo de texto é feita. A anotação de texto é estritamente associada com a extração de conhecimento útil por causa do uso de classificadores e marcas para etiquetar e rotular o texto.

As novas aplicações de anotação de texto permitem que a tecnologia de mineração de dados substitua a interpretação de texto tradicional devido ao uso de *corpora* na fase de análise.[6]

2.1.3 Diferença entre Subjetividade e Objetividade (2003-2005)

No início do ano 2000, época que marca o início da terceira fase do friso cronológico definido, a diferença entre subjetividade e objetividade foi a principal discussão que produziu maior classificação de polaridades de maneira correta. [7]

(Stoyanov *et al.*, 2003) propuseram uma abordagem para extrair informação com o intuito de categorizá-la e organizar opiniões em texto naturalmente ocorrido, de forma a suportar meios de resposta a questões multi-perspetiva, ou a tarefas de resposta a questões que requerem habilidade de pesquisa de opiniões no mesmo texto. Assim, apresentava-se

[6] Em linguística, um *corpus* (*corpora plural*) ou corpus do texto é um conjunto alargado e estruturado de textos (hoje em dia eletronicamente armazenado e processado). São usados para a execução de processos de análise estatística e de teste de hipóteses, permitindo verificar ocorrências ou validando regras linguísticas dentro de um território específico de linguagem.
[7] A análise de polaridade tem em conta a quantidade de termos considerados positivos ou negativos e que constituem parte da frase analisada.

um esquema de baixo nível para anotar e identificar avaliações, emoções e especulações na linguagem e construir um *corpus* do documento, bem como uma apresentação possível de um sumário para a codificação da coleção de opiniões expressada num documento. A classificação a diferentes níveis, tanto ao nível documental como ao nível frásico, permitiu obter tanto resultados sumários como detalhados, respetivamente. O desenvolvimento de abordagens na terceira fase acelerou a mineração de opiniões para uma nova era aplicacional e colocou bem no cerne da *Internet* a análise de texto não estruturado em vários domínios e contextos, a denominada mineração de dados *Web*.

2.1.4 Aplicações Web de Mineração de Dados (2006-2007)

Ao começar com a investigação de *Bing Liu*, o ano de 2006, marca o nascimento da mineração de opiniões. Nesses anos (Eguchi and Lavrenko, 2006) propuseram modelos de recuperação sentimental, usando uma *framework* de modelação de linguagem generativa e assumindo que a polaridade de interesse de sentimentos é definida pelo utilizador com a consideração de dependência de tópicos. Nesta quarta fase assistiu-se ao crescimento do nível aplicacional do campo em estudo. Tendo aparecido uma enorme diversidade de aplicações no domínio. Política, saúde, entretenimento, finanças, marketing, e *social networking* foram algumas das áreas em que emergiram várias aplicações foram implementadas, envolvendo aspetos de mineração de dados *Web* ou de mineração de opiniões na *Web*. Tudo isto foi realizado para providenciar auxílio na análise de feedbacks e

comentários de utilizadores nos múltiplos sites de expressão social, de forma a que fosse possível, posteriormente, fazer a avaliação de produtos ou de serviços. Após a interpretação de textos, da anotação de textos, e a evolução das várias técnicas de classificação, a mineração de opiniões começou a ser utilizada em diferentes domínios aplicacionais para identificar e classificar texto não estruturado.

2.1.5 Recursos Léxicos (2008-2011)

Em (Liu, Zhai and Pedrycz, 2012) foi concluído que o agrupamento de expressões características constituía um fator crítico para um sumário efetivo das opiniões que reduz o tempo consumido no seu agrupamento em categorias de características. Por outro lado (Valarmathi and Palanisamy, 2011) investigaram a classificação de mineração de opiniões baseando-se em palavras de corpus que são comummente usadas em *reviews*.

Nesses trabalhos, a aprendizagem supervisionada foi usada para classificação com exemplos etiquetados para o atributo das classes. A medida de precisão da classificação depende do modelo construído usando os dados históricos que, precisamente, preveem a etiqueta de exemplos não etiquetados. Assim, esta última fase da cronologia retrata uma abordagem comum para a mineração de opiniões como uma abordagem léxica. Recursos léxicos tais como o *WordNet* e o *SentiWordNet* (Valarmathi and Palanisamy, 2011) providenciam partes do discurso agrupado em formato especial conhecido como o agrupamento cognitivo.

Atualmente, continuam a ser desenvolvidas aplicações nas áreas que retiram proveito deste campo de estudo, continuando a serem realizados trabalhos de investigação para descobrir mais e melhores formas de otimizar a performance de todos os processos envolvidos num sistema de mineração de opiniões e de análise de sentimentos.

3 Análise de Opiniões e Sentimentos

A análise/mineração de opiniões e de sentimentos é a área de pesquisa e investigação que procede à implementação de métodos cujo objetivo reside no sentido de realizar, executar e implementar sistemas de aprendizagem automáticos com o objetivo de captar, e, sobretudo, interpretar opiniões e sentimentos humanos, tendo como base um conjunto de textos escritos em linguagem natural (Bhuiyan, Xu and Jøsang, 2009). Além de adotar um perfil tecnicamente desafiante, a técnica de mineração de opiniões e sentimentos enfrenta, a vários níveis, dificuldades bastante particulares especialmente devido a várias tarefas incluídas em todo o processo, como por exemplo o processamento de linguagem natural. Em adição, existem mais alguns desafios no que à mineração de opiniões e de sentimentos concerne. Não obstante tal facto, a mineração de opiniões e de sentimentos apresenta-se como um domínio de estudo repleto de proveitos e possibilidades de aplicação. Exemplo disso é o caso das empresas que procuram constantemente a melhoria de desempenho em relação ao serviço que prestam aos seus clientes. No seguimento desta ideia impõe-se, naturalmente, uma questão: "Como ordenar e estabelecer uma lógica, corretamente estruturada, para avaliar o desempenho da empresa em questão juntos dos seus clientes?". Adicionalmente, é comum também pretender-se saber "Como definir uma lógica bem arquitetada para avaliar a satisfação dos clientes com os serviços da empresa/instituição em causa?". Na tentativa de redarguir tais questões formuladas, é necessário desenvolver esforços para fazer a implementação de soluções, como é o caso de aplicações para o estabelecimento de índices de satisfação.

As opiniões e sentimentos podem constituir, assim, a base para o estabelecimento de um conjunto de índices de satisfação, que atuam como métricas de avaliação para possibilitar o estudo de perspetivas de análise,

devendo ser tomadas em conta por parte de quem pretender medir o seu desempenho na área em que atua (Archak, Ghose and Ipeirotis, 2011).

Deste modo, a análise de opiniões e de sentimentos revela-se de grande utilidade no quotidiano empresarial, uma vez que as empresas necessitam fazer a interpretação dos dados que vão obtendo, resultantes das suas vendas no mercado em que atuam. Assim sendo, as empresas necessitam de vários recursos para compreender e conseguir "interpretar" o mercado, saber como ele se comporta de acordo com as vendas dos seus produtos ou de outros aspetos subjacentes. Usando técnicas de análise de opiniões e de sentimentos é possível "ler" o futuro do mercado em que estão a agir, o que permitirá as justificações para melhores tomadas de decisão empresarial.

A análise de sentimentos, também denominada mineração de opiniões, é o campo de estudo que analisa as opiniões, sentimentos, avaliações, apreciações, atitudes e emoções das pessoas perante elementos de dados sobre produtos, serviços, organizações, indivíduos, problemas, eventos, tópicos e seus respetivos atributos, sendo um contexto de problema bastante extenso e transversal a diferentes domínios (Liu, 2012). O termo *"Análise de Sentimentos"* terá sido primeiramente utilizado por (Nasukawa and Yi, 2003), enquanto que o termo "Mineração de Opiniões" teve a sua primeira ocorrência em (Dave *et al.*, 2003). Todavia, o interesse académico pelo estudo dos sentimentos e das opiniões foi crescendo bastante desde o ano 1997, sensivelmente, existindo alguns trabalhos relevantes neste campo de estudo – veja-se, por exemplo os trabalhos de (Das and Chen, 2001), (Hatzivassiloglou e McKeown, 1997), (Morinaga *et al.*, 2002), (Pang, Lee e Vaithyanathan, 2002), (Peter D. Turney, 2001) ou (Wiebe, 1990).

Apesar dos campos linguístico e de processamento de linguagem natural (PLN) serem já bastantes extensos, sobre o processamento de opiniões e de sentimentos, foi realizada, em comparação com o momento atual, uma quantidade de estudos significativamente menor até ao ano 2000. A partir desse mesmo ano, este campo tornou-se numa área de investigação muito ativa. As razões para este crescimento de trabalhos de investigação assentam,

essencialmente, na procura que os sistemas capazes de incluir em si a análise de sentimentos e de mineração de opiniões despertaram. Pelas referências apresentadas, podemos ver que o ano 2000 acolheu um aumento significativo dos desenvolvimentos ocorridos neste campo de estudo. Contudo, antes do ano de 2001, e antes de todo o interesse existente atualmente no campo da análise de opiniões e mineração de sentimentos, existiam já vários estudos relativos à análise e interpretação de metáforas, de adjetivos sentimentais, da subjetividade, dos pontos de vista e dos afetos (McKeown *et al.*, 1999), (Knight and Hatzivassiloglou, 1995), (Smadja, McKeown and Hatzivassiloglou, 1996), (Wiebe, 2000), (Hatzivassiloglou and Wiebe, 2000).

As opiniões são o elemento chave de muitas tomadas de decisão, desde as mais básicas até às mais complexas. A influência de certas opiniões na compra (ou não) de certos produtos desempenha um fator precípuo sobre as massas do conjunto de pessoas designadas por possíveis clientes de uma empresa.

Tomando, por exemplo, o caso do *Youtube*, e do que este possibilita no que toca à influência de massas, assiste-se hoje em dia à existência de canais específicos nessa mesma plataforma, cujo conteúdo é produzido por indivíduos que possuem por objetivo a crítica de produtos existentes no mercado.

De facto, existe uma enorme quantidade deste tipo de canais, e, consequentemente, vídeos, nos quais onde são executadas várias ações tais como analisar, e executar *críticas* sobre produtos à venda no mercado, ou de serviços disponibilizados por empresas. As críticas feitas pelos canais vão desde vídeos que criticam negativamente o preço que uma dada marca fixa para determinada reparação de um ecrã de um *smartphone* que comercializa, até a vídeos que analisam a potência, os modos de condução, o isolamento de ruído, entre outras características, de um automóvel comercializado no mercado. Na maior parte das vezes, estes vídeos pertencem a canais já com um número significativo de subscritores, que, maioritariamente, comentam e atuam de forma ativa no canal. Além disso, algumas vezes, os produtos em análise nos vídeos ainda nem se encontram no mercado ou então são analisados numa fase ainda

bastante inicial da vida do produto no mercado. Isto quer dizer que, as opiniões e os sentimentos revelados em relação a determinado produto ou serviço podem começar a ser expressos ainda mesmo antes de o produto atingir o mercado. O peso influenciador que as várias plataformas de ação e partilha social, como o caso do *Youtube*, têm grande influência, obrigando as próprias marcas, muitas das vezes, a reinventar-se e a apostar mais na publicidade e na abertura de portas através de vídeos destinados aos utilizadores da *Web*. Todos estes fatores obrigam frequentemente a um reposicionamento das marcas, empresas e indústrias em relação a como apresentam um dado produto e, principalmente, a restringir e disciplinar de forma mais efetiva as normas de atendimento a clientes. Assim, cada cliente desempenha um papel preponderante na ação de uma empresa, uma vez que a partilha de experiências sobre algo comercializado, hoje em dia, é muito maior, e tem, de facto, grande impacto. Isto significa que, o que para uma marca pode parecer e ser apenas um "simples" cliente e "apenas mais um cliente", pode rapidamente tornar-se no cliente autor de um vídeo em que, no caso de incorreto atendimento ou incorreta prestação de serviços, expõe o seu relato negativo com a empresa numa dada plataforma social, levando a que muitas mais pessoas tenham conhecimento do sucedido, o que danifica, naturalmente, a imagem da empresa em questão.

Atualmente assiste-se, não raras vezes, a tomadas de ação por parte de empresas para com sujeitos individuais, pois estes manifestaram em plataformas socias o seu descontentamento. para com o serviço que obtiveram por parte dessa mesma empresa, tendo o seu manifesto atingido um grande número de "*shares*" o que levou a uma danificação da imagem da empresa.

Por vezes, tais manifestos atingem um tão elevado número de "*shares*" e uma exposição mediática grande quando a *review* atinge os *media*. Como reação a este tipo de coisas, notam-se atitudes responsivas por parte das empresas quando esses momentos são atingidos como, por exemplo, uma reparação gratuita do ecrã de um *smartphone* pelo qual pediam um valor excessivo, ou até

a correta reparação do *smartphone*, uma vez que o repararam o mesmo, mas utilizando peças não originais da marca.

Hoje, são inúmeros os assuntos sobre os quais um utilizador pode expressar a sua opinião nas redes sociais. Exemplos de tal são uma eleição política que gera discussão nas redes sociais, a avaliação sobre um produto, a crítica sobre um evento desportivo, entre outros. As aplicações industriais de sistemas que possuem incorporados em si a capacidade de analisar sentimentos e minerar opiniões, retirando proveito da informação que destas tarefas provém, são bastantes. Tal estimulou bastante o progresso nos estudos sobre a interpretação de opiniões e de sentimentos, na tentativa de se obter conhecimento útil sobre as críticas feitas a produtos ou a serviços de um dado grupo empresarial ou de qualquer conjunto de entidades que atue no mercado e pretenda possuir algum tipo de feedback sobre os produtos e serviços que comercializa.

A geração de dados desencadeada pela ação de mercado de uma dada empresa acaba por poder ser medida, e de forma significativa, ao providenciar suporte (Cambria *et al.*, 2013) a um sistema que capta informação nessas mesmas fontes de dados (*críticas*, discussões, comentários, opiniões ou sentimento, etc.) (Hu and Liu, 2004b), e que permite que seja retirada uma enorme quantidade de conhecimento útil dos elementos que usualmente produzem. Neste casos, todo e qualquer tipo de comentário é analisado (Bhadane, Dalal and Doshi, 2015), numa primeira fase, pelo sistema, sendo avaliado em vários parâmetros, como por exemplo as críticas geradas num serviço de uma empresa que presta suporte ao cliente da mesma. Depois disso, surge o desafio de implementar índices com base nos dados resultantes de todo o processo de análise, para que, desta forma, se possa fazer uma análise mais precisa e axiomática dos resultados obtidos – em (McGlohon, Glance and Reiter, 2010) está apresentado um caso de estudo que analisa o facto de, sobre um mesmo produto, existirem inúmeras críticas, em múltiplos sites, no qual foram efetuadas várias tentativas para solucionar a heterogeneidade existente de classificações sobre um mesmo produto.

Desde modo, a organização dos dados com base em índices pode proporcionar vários benefícios, tais como a rápida execução intuitiva de estudos sobre os dados interpretados pelo sistema, de forma a se poder aumentar a facilidade de retirada de ilações sobre grandes volumes de dados.

Concluindo, é de grande utilidade prática a utilização de dados como os referidos para o desenvolvimento de estudos sobre produtos, sistemas ou serviços que uma empresa vai lançando no mercado e, especialmente, quando esta necessita de adquirir algum tipo de conhecimento útil sobre a opinião dos seus clientes acerca dos seus produtos, sistemas ou serviços.

3.1 Níveis de Análise

Na mineração de opiniões e sentimentos podem definir-se diferentes graus de granularidade, nomeadamente os três que a seguir se apresentam: nível documental, nível frásico e nível de aspetos e entidades. Vejamos em que consiste cada um destes níveis de análise.

Ao nível documental a tarefa é classificar toda uma opinião expressa num documento, categorizando-a em sentimentos negativos ou positivos. Como exemplo disto, vejamos o caso de uma crítica relativa a um produto, em que o sistema determina se a crítica expressa revela uma opinião positiva ou negativa sobre o produto em si. Esta mesma tarefa é vulgarmente conhecida como classificação de sentimentos ao nível documental. Este nível de análise assume que cada documento expressa opiniões numa entidade singular, ou seja, num produto apenas. Assim, neste nível, não são aplicáveis documentos que avaliem em simultâneo mais do que uma entidade.

Ao nível frásico pretende-se determinar se uma frase revela um sentimento positivo, neutro ou negativo. Entende-se por opinião neutra uma opinião resultante de uma frase que não expressa qualquer tipo de opinião. Este nível de análise é intrínseco à classificação de subjetividade (Wiebe, Bruce e O'Hara, 1999), que efetua a distinção entre frases que expressam informações fatuais, denominadas de frases objetivas, e frases que expressam perspetivas e opiniões subjetivas. Todavia, deve ser notado que a subjetividade não é equivalente à presença de sentimentos, uma vez que mesmo as frases objetivas podem implicar a presença de opiniões, como acontece, por exemplo, na frase "*Comprei o telemóvel ontem e o ecrã avariou.*"

As análises referentes ao nível documental e ao nível frásico não revelam com exatidão os verdadeiros sentimentos expressos pelas pessoas. Se aumentarmos significativamente a granularidade em relação à análise a ser efetuada, realizaremos a denominada análise ao nível dos aspetos e entidades propriamente ditas. O nível dos aspetos começou por se denominar de nível das características (como em "*sumarização e mineração de opiniões baseadas em características*" (Hu and Liu, 2004a)). Ao invés de analisar a forma como se dá a construção da frase, utilizando a linguagem (documentos, parágrafos, frases ou orações), o nível das características afetas aos produtos ou serviços tem em conta, diretamente, a opinião em si propriamente dita. Uma opinião consiste num sentimento (positivo, neutro ou negativo) e num desígnio da opinião sobre algo que se encontra sob avaliação. Uma opinião sem um objetivo ou sem uma finalidade identificadas não se revela de grande utilidade, uma vez que o propósito da análise é ter em conta a avaliação expressa pela opinião que se encontra a ser analisada. A importância dos propósitos das opiniões é o cerne a partir do qual se compreender melhor o problema da análise de sentimentos. Vejamos a seguinte frase: "*Embora o serviço não tenha sido prestado da melhor forma, adoro na mesma este cruzeiro.*". O tom patente nesta frase parece ser positivo. Não obstante tal facto, é de notar de igual modo que a frase não é, de

todo, inteiramente positiva. Sobre o cruzeiro como um todo, no seu âmbito geral, é enunciada um *review* positiva. Todavia é expresso, também, um sentimento de desagrado para o serviço prestado nesse mesmo cruzeiro.

Em muitas aplicações, os desígnios das opiniões são descritos pelas entidades ou pelos seus diferentes aspetos. Apesar de tal, o objetivo deste tipo de análise é descobrir sentimentos que se encontram patentes em entidades e/ou nos seus aspetos. Numa frase como "*A qualidade da câmara do computador é alta, mas a sua autonomia da sua bateria é relativamente baixa.*", são avaliados dois aspetos: a qualidade da câmara e a autonomia da bateria desse mesmo computador, que neste caso, em específico, é a entidade em análise. O sentimento revelado em relação à qualidade da câmara do computador é evidentemente positivo, mas o sentimento expresso relativamente à autonomia da bateria é claramente negativo. As características do computador, nomeadamente a câmara em relação à sua qualidade e definição e a bateria em relação à sua autonomia, são os alvos desta opinião expressa através da frase apresentada como exemplo.

Baseado neste tipo de análise, um sumário estruturado de opiniões sobre entidades e os seus aspetos respetivos pode ser realizado, tornando, desta forma, texto não estruturado em conhecimento útil devidamente estruturado, que pode ser usado para todos os tipos de análise qualitativas e quantitativas. Ambos os tipos de classificação analítica, nomeadamente a nível documental e a nível frásico, são problemas desafiantes. Ao nível da análise dos aspetos o problema possui maior nível de complexidade.

Em (Jindal and Liu, 2006) são reveladas dois outros tipos de opiniões: usuais e comparativas. Uma expressão usual expressa um sentimento apenas sobre uma entidade particular ou sobre o aspeto de uma entidade, como por exemplo na frase "Esta bebida sabe muito bem.". Esta frase expressa um sentimento positivo no gosto positivo de certa bebida. Por sua vez, uma opinião comparativa procede à comparação entre múltiplas entidades baseada em alguns dos seus aspetos partilhados, como se nota, por exemplo, na frase "*A Apple é*

melhor que a Samsung.", na qual se comparam duas empresas que atuam em várias mercados de diferentes domínios, mas que, neste caso específico, se analisa e compara o desempenho das duas no mercado dos telemóveis, o que expressa uma preferência pela primeira ao invés da segunda, respetivamente.

3.2 Léxico Sentimental

A maior parte dos indicadores de elementos sentimentais são exatamente as palavras sentimentais presentes nas frases submetidas à análise, que também são denominadas de palavras de opinião. Estas são palavras comummente utilizadas para a expressão de sentimentos positivos ou negativos. *Bom*, *fantástico*, *espetacular*, *utópico*, e *quimérico*, são alguns exemplos de palavras sentimentais positivas, enquanto que as palavras *mau*, *pobre*, e *terrível* são palavras sentimentais obviamente negativas. Aparte de tais conjuntos de palavras, note-se ainda a existência de expressões especificamente utilizadas no contexto de um idioma, como é o caso de frases como "*O automóvel custou-me os olhos da cara.*" ou "*Estava a chover a cântaros e o recinto onde o evento se realizou não era coberto.*". As expressões "custar os olhos da cara" e "chover a cântaros" significam, respetivamente, o preço elevado de algo, e uma grande quantidade de precipitação, respetivamente, sendo que se revelam de grande especificidade na Língua Portuguesa - na Língua Inglesa existe também, a título de exemplo a expressão "*Raining cats and dogs.*" que significa uma grande quantidade de chuva. A lista na qual se incluem estas palavras e expressões é denominada de léxico sentimental ou léxico opinativo. Ao longo do tempo, e da linha cronológica do estudo que respeita este campo, os investigadores investigaram e procuraram implementar algoritmos que compilassem tais

conjuntos léxicos. Embora as palavras sentimentais e as frases desempenhem um papel primordial na análise de sentimentos, estes elementos não são suficientes para fazer a análise de sentimentos. Aqui podemos encontrar variadíssimos problemas são enfrentados tais como:

1. Uma palavra ou uma expressão sentimental positiva ou negativa pode possuir orientações opostas em diferentes domínios aplicacionais.

   A frase "*As colunas que adquiri conseguem emitir som a um nível bastante elevado.*", revela um sentimento positivo para com a performance das colunas adquiridas. Pelo contrário na frase "A máquina de lavar a roupa que adquiri emite um som a um nível bastante elevado quando se encontra em funcionamento.", revela um sentimento negativo pois a máquina de lavar adquirida emite bastante ruído quando se encontra em funcionamento.

2. Uma frase que contém palavras sentimentais pode não expressar sentimento algum. Tal acontece frequentemente em vários tipos de frases. Dois tipos importantes de frases são as frases interrogativas e condicionais, como por exemplo, "*Pode, por favor, dizer-me quais os telemóveis que são bons?*" ou "*Se vir um automóvel bom no stand compro-o.*". Ambas as frases apresentadas anteriormente contêm palavras sentimentais tais como "*bons*" e "*bom*", mas, apesar disso, nenhuma expressa uma opinião positiva ou negativa em nenhum dos objetos supracitados. Apesar de tal facto, nem todas as frases condicionais ou interrogativas estão isentas de expressar sentimentos tal como acontece na frase "*Quanta custa a reparação deste automóvel terrível que está sempre com problemas a todo o momento?*" e "*Se procuras um bom telemóvel compra um iPhone.*"

Estas frases são, claramente, uma frase interrogativa e uma frase condicional que expressam sentimentos bem claros no que toca a certos produtos em questão, respetivamente.

3. Frases que fazem uso do sarcasmo com ou sem presença de palavras sentimentais são bastante complexas de analisar. "*Que telemóvel fantástico! Avariou-se passados dois dias após a compra do mesmo!*" é um exemplo de frase sarcástica que complica significativamente a tarefa de analisar os sentimentos nela presentes nela mesma. Em vários tipos de críticas, o sarcasmo marca presença como forma de expressar maioritariamente opiniões negativas. Geralmente, nota-se uma grande quantidade de opiniões que, com o sarcasmo, pretendem exagerar a crítica negativa e chamar a atenção de um produto ou de um serviço que se encontra sob avaliação. Os domínios nos quais existe a utilização de sarcasmo nas críticas são vários, estando entre eles as críticas de produtos, de serviços, ou de discussões políticas, entre outras.

4. Um grande número de frases sem palavras reveladoras de sentimentos também implicam opiniões nelas patentes. Muitas dessas frases são objetivas e são usadas para expressar alguma informação factual. Existem vários exemplos de frases deste tipo como, por exemplo, "*Este tipo de condução gasta os jogos de pneus muito rápido.*". Nesta frase existe uma crítica negativa quanto à forma de condução do carro em questão. Já a frase "*Após três dias de uso, o ecrã do computador começou a revelar problemas de delaminação.*", que revela uma opinião negativa sobre o ecrã do computador, que após um curto espaço de tempo desde a sua compra, começou a apresentar sinais de delaminação.

5. Nos dias de hoje, além dos problemas acima explanados, existem vários outros fatores que dificultam a tarefa de análise de sentimentos nos dados textuais que um dado sistema processa. Os denominados *emojis* como por exemplo: ☺ ou ☹, são elementos importantes que são fundamentalmente utilizados para melhorar a análise e compreender a opinião que se encontra a ser expressa pelo utilizador. Verifica-se, atualmente, uma grande utilização de *emojis*, como forma de expressar melhor o sentimento que se pretender implantar no texto, tornando a opinião mais forte, e mais suportada graficamente. O facto de existirem categorias pelas quais os emojis são organizados permite, desde logo, retirar ilações sobre o tipo de domínio ao qual a opinião se refere. A título de exemplo veja-se a frase seguinte na Figura 4.

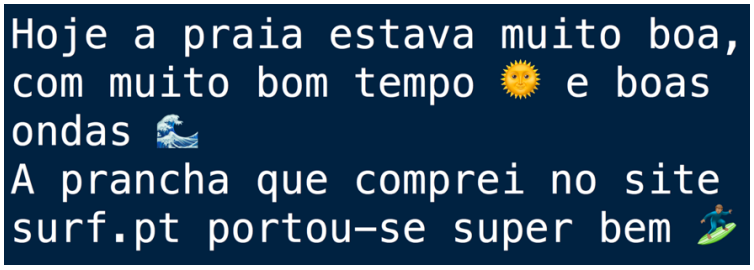


Figura 4: Exemplo de crítica de ecommerce com recurso a emojis

Na Figura 4 é possível observar uma *review* feita num domínio de compras *online*, mais especificamente, no domínio do mercado de pranchas de *surf*. Na crítica apresentada o utilizador afirma estar feliz com a sua compra, uma vez que a prancha adquirida revelou grande performance e desempenho aquando da sua utilização. Note-se que o primeiro *emoji*, 🌞, revela, desde logo, uma referência meteorológica, e, de facto, precede a frase que afirma que esteve bom tempo na praia e que esta se encontrava em condições agradáveis com boas ondas - esta última informação é reforçada pelo emoji 🌊. Quanto à última frase, esta

refere-se especificamente à qualidade da prancha adquirida no *site surf.pt*, sendo importante destacar aqui o sentimento positivo que qualifica o desempenho e a performance da prancha nas ondas.

Além da informação textual, uma vez mais, é feita a "hipérbole" da afirmação anterior através do emoji 🏄 que faz uma referência clara à atividade desportiva denominada de *surf*.

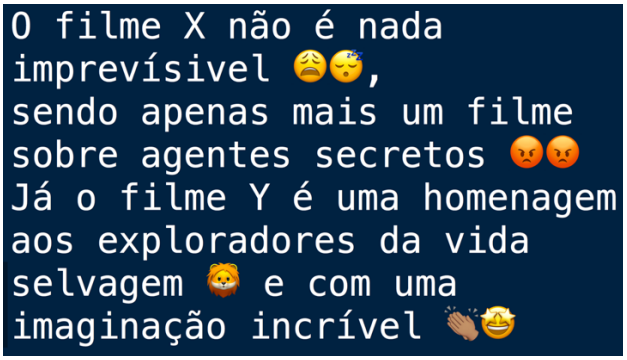


Figura 5: Exemplo de crítica de cinema com recurso a emojis

Por sua vez, na Figura 5 pode ver-se um exemplo de uma crítica de cinema recorrendo ao uso de emojis. Neste exemplo, é possível notar que os dois primeiros pares de emojis (😩😴 e 😡😡) denotam um sentimento negativo em relação ao filme em causa (filme X), sendo afirmado que o filme é uma repetição de outros, e que não cativou o utilizador em questão pela falta de surpresas e regularidade do guião, quando comparado com outros filmes. O primeiro par de *emojis*, 😩😴 denota grande aborrecimento em relação à visualização da película, enquanto que o segundo par de *emojis*, 😡😡 denota um sentimento verdadeiramente negativo no que toca à avaliação do filme por parte utilizador. Estes são, indubitavelmente, dois pares de *emojis* que sustentam e provocam um aumento na robustez da opinião efetuada.

Na segunda afirmação da Figura 5 é possível observar que o filme Y provocou uma reação positiva no utilizador, que recorreu a um *emoji*, nomeadamente o  para, não só, explanar o domínio em que a crítica é realizada, como também para reforçar o contexto do filme, que é referido como um filme sobre a exploração da vida selvagem. Os dois últimos *emojis* da segunda *review* -  - reforçam o sentimento positivo que é tido como conclusão da análise textual do texto que os antecede.

O domínio dos emojis é, certamente, um domínio que pode auxiliar a análise dos sentimentos presentes num texto que se pretenda analisar. Porém, é também necessário ter bem presente que, em muitos casos, os emojis são utilizados também para ironizar a opinião. Em casos extremos, chegam a ser utilizados emojis que não se correlacionam com a opinião que está a ser formulada, num tipo de utilização de baixo ou médio nível de complexidade de compreensão. Grande parte das vezes dá-se a ocorrência deste tipo de situações anteriores, verificando-se um elevado nível de complexidade no que toca a compreender bem a intenção e a polaridade dos sentimentos presentes na opinião em causa, o que dificulta, obviamente, a mineração de opiniões e a análise de sentimentos sobre o texto em análise.

Aliado ao campo dos emojis, existe também o campo das hashtags. As hashtags são elementos que podem estar presentes nas críticas, tendo já obtido uma difusão bastante elevada em plataformas tais como o *Twitter*, *Instagram*, ou o *Facebook*, entre outras. As *hashtags* podem ser vistas como *keywords*, palavras-chave do contexto em questão, quer este seja um tópico de discussão, um evento de qualquer tipo, ou todo e qualquer elemento do universo em contexto com a opinião ou texto escrito. A constituição de uma *hashtag* faz-se pela presença do símbolo *"#"* seguido do elemento a que se quer referir a *hashtag*. Em muitas das plataformas que permite a existência de *hashtags*, estas são *links* para o conjunto de frases ou imagens, entre outras que fazem parte do conjunto de conteúdos conotado e identificado com a mesma *hashtag*. A maioria

das plataformas permite ainda, pesquisar por *hashtag* e até seguir novas publicações que venham a ser feitas com a *hashtag* em causa. As *hashtags* foram idealizadas como um modo de indexar e assim, organizar o conteúdo publicado na *Web*. Uma analogia passível de ser feita é a comparação com um dicionário em *Python*: as *hashtags* funcionam como as *keys* do dicionário que representa o universo das *hashtags*, ao passo que todo o conteúdo ao qual é anexado, usualmente como descrição, uma *hashtags* passa a ser incluída no conjunto de elementos que correspondem ao valor dessa mesma *key*. Sendo assim, a organização criada pelas *hashtags* pode ser vista como um mapeamento par-valor, que é, fundamentalmente, a base que está por trás da indexação de *hashtags*, permitindo assim a organização e eficiência quando se quer pesquisar um dado conteúdo relativo a uma *hashtag* em específico. Vejamos, como exemplo de aplicação de *hashtags*, o seguinte caso: "*Ontem vi um jogo bastante disputado integrante da competição da Liga dos Campeões. #ChampionsLeague #Football".*

Neste caso, é possível observar que, além de uma crítica sobre um jogo da Liga dos Campeões sobre o qual é dito que foi bastante disputado, são utilizados duas outras *hashtags*: *#ChampionsLeague #Football*. Estas hashtags pretendem adicionar esta frase ao arquivo de frases contidas no conjunto indexado pelo *hashtag ChampionsLeague* e pelo *hashtag Football*.

Além de indexarem conteúdos por assunto, as hashtags são também elementos no que toca a criar correntes relativamente a causas ou a eventos. Estas são também capazes de aumentar a visualização de certas marcas e de colocar em expansão os seus produtos e serviços, através da mediatização de certas hashtags correlacionadas com produtos e serviços em específico. Várias das hashtags que despertam a atenção para eventos de cariz social, cultural, desportivo, entre muitos outros. Vários exemplos de *hashtags* foram utilizados até à data presente possuindo um propósito específico, como por exemplo na frase "*#ootd: Outfit Of The Day"*, que serve o propósito de, para todo e qualquer tipo de *blogger* ou utilizador, promover o *outfit* que usa na foto que publica, por

exemplo no *Instagram*. Existem muitos mais exemplos deste tipo de *hashtags*. Vejamos mais três outros exempos:

1. *#BlackLivesMatter*, que é utilizada aquando de vários conteúdos que apoiam a comunidade negra em variados casos de desigualdade e injustiça social.

2. *#NowPlaying* que é frequentemente utilizado para descrever conteúdos como vídeos que demonstram uma performance ao vivo naquele preciso momento em que é feita a publicação desse mesmo conteúdo na plataforma em questão.

3. *#IceBucketChallenge*, uma *hashtag* muito popular, que é utilizada para indexar conteúdo produzido com o objetivo de angariar fundos para a associação americana de esclerose lateral amiotrófica.

3.3 Opiniões e Sentimentos

A matéria-prima de um sistema de análise de dados, como o que se encontra em estudo neste trabalho, são os dois elementos já tão referidos nesta dissertação, nomeadamente opiniões e sentimentos (Abbasi, 2010). Com o objetivo de compreender que opiniões e sentimentos estão a ser expressados e que se encontram patentes na informação textual que está a ser analisada, costumamos colocar um conjunto de questões bastante concreto,

nomeadamente "Quem?", "O quê", "Onde?", "Quanto?", entre outras. (Breck, Choi and Cardie, 2007)(Chen, 2012)

Durante os últimos anos, a investigação e a pesquisa neste campo tem aumentado de forma muito significativa, devido a várias razões. Entre elas, o grande número de aplicações e de desafios científicos que as matérias relacionadas com este campo têm imposto (Balazs and Velásquez, 2016)(Xia *et al.*, 2007). Estas duas peças, existentes em abundância hoje em vários tipos de plataformas, desempenham um papel altamente relevante no fundamento e construção de sistemas de análise de dados, que varia, obviamente, consoante o propósito de cada um. Uma opinião define-se como uma perspetiva, atitude, emoção ou avaliação tendo em conta algo que representa o objeto da opinião. Toda a opinião possui uma polaridade, nomeadamente: positiva, negativa ou neutra, ou seja, isto é uma característica da opinião (Jindal and Liu, 2008)(Khan *et al.*, 2009). A opinião pode ainda segmentar-se em dois grupos: direta ou comparativa, se avalia diretamente algo ou se efetua comparações entre mais do que um objeto de opinião, respetivamente. Bing e Liu (Liu e Bing, 2010) categorizaram também a informação textual existente em factos e opiniões, afirmando que os factos podem ser caracterizados como expressões objetivas sobre entidades, eventos e suas propriedades, e definindo opiniões como expressões que contêm um carácter usualmente subjetivo e que descrevem os sentimentos, apreciações ou emoções das pessoas.

Muitos dos trabalhos desenvolvidos sobre a informação textual existente possuem especial enfoque na mineração, análise e retirada de ilações, sobre a forma de informação factual, como, por exemplo, pesquisas *Web*, classificação de texto, *clustering* de texto, entre outras tarefas de processamento de linguagem natural e de mineração de texto.

O processamento de opiniões é um objeto de estudo algo recente, até porque, num passado também recente, as plataformas de *críticas* se revelaram bastante úteis para aqueles que pretendem obter conhecimento sobre algo que

já foi objeto de opinião de outros indivíduos, com o intuito de concluir se avançam para a compra de certo produto ou obtenção de certo serviço.

Fundamentalmente verifica-se que um indivíduo que necessite de tomar uma decisão, pretende ouvir a opinião de outros que já tenham experienciado o objeto da decisão a tomar e que assim possam partilhar os prós e os contras da sua experiência. Tal verifica-se não só para decisões a nível individual, mais experienciadas a nível do cliente, como para decisões organizacionais ou empresariais, personificadas pelas empresas atuantes no mercado.

O aparecimento da *Web* veio aumentar extensamente os volumes de texto de opinião existentes, substituindo os mais primitivos meios de obter conhecimento sobre a experiência de um produto ou de um serviço.

Tanto para o indivíduo que pretender obter informação sobre a experiência de outros utilizadores acerca de um produto ou de um serviço, como para a empresa que pretender conhecer as opiniões dos seus clientes sobre os produtos que vende ou os serviços que presta, o nível de texto de opinião existente é benéfico. Os indivíduos podem informar-se melhor antes de efetuar uma compra ou obter um serviço e as empresas podem analisar de forma específica as opiniões existentes sobre os seus produtos e serviços que comercializam, providenciando assim uma fundamentação sustentada e originando uma boa margem de manobra em relação à melhoria dos seus produtos e serviços. Não obstante tudo o que acima foi afirmado, encontrar fontes de opiniões credíveis e monitorizá-las na *Web* pode constituir uma tarefa morosa e bastante intrincada. Tal deve-se ao extenso número de diversas fontes de opiniões e às especiais características deste tipo de texto, como são o conteúdo baseados em emoções e sentimentos. A maior parte das vezes as opiniões encontram-se expostas em longas publicações em fóruns e em blogs, sendo de elevado grau de dificuldade a extração, leitura, sumarização e posterior disposição destas através de, a título de exemplo, índices de satisfação, o objeto de estudo do presente trabalho. Assim, são necessários sistemas que automatizem a descoberta de textos que contêm cargas relativas a opiniões e que, posteriormente, as sumariem para

facilitar a sua análise e estudo. A análise de sentimentos surge como a resposta a estas necessidades, sendo um problema de processamento de linguagem natural ou mineração de texto.

Em virtude do enorme valor contido na análise de sentimentos e do enorme leque de aplicações práticas que podem beneficiar de sistemas com tal capacidade, os processos de investigação sobre este campo de estudo, tanto a nível académico como a nível industrial, aumentaram muito. A existência de sistemas munidos com este tipo de análise é uma realidade reconhecida a nível mundial por várias empresas presentes em diferentes tipos de mercado. Isto deve-se à utilidade desta técnica, que é transversal a todos os tipos de mercado, dado que basta existir a intenção ou a necessidade de conhecer os pontos de vista de um dado público-alvo sobre um conjunto de produtos que são comercializados e que carecem de melhorias ou sobre um qualquer serviço prestado por determinada empresa. Qualquer tipo de comércio que suscite o conhecimento da opinião dos clientes beneficia em muitos fatores com a presença sustentada de sistemas que fazem uso da análise de opiniões e de sentimentos expostos nas opiniões de quem usufrui do produto ou do serviço respetivamente vendido e/ou prestado.

3.4 Polaridade dos dados

A polaridade dos dados é inerente à subjetividade remetida para o tópico das opiniões, sendo o elemento que faz a opinião carregar a expressividade a si afeta. Analisar sentimentos e minerar opiniões são ambas tarefas que se implementam através de processos de deteção da polaridade contextual do texto em análise (Devika, Sunitha and Ganesh, 2016). A própria distinção de uma crítica em *"thumbs up"* e *"thumbs down"* ou em *"like"* e *"dislike"* caracteriza um exemplo de classificação de polaridade dos dados (Cambria *et al.*, 2013).

A existência de polaridade do texto (Taboada *et al.*, 2011) está inteiramente relacionada com o campo das emoções, ou seja, com os sentimentos, os pensamentos, as perspetivas, ou as atitudes subjetivas que estão contidos no texto. As emoções são objeto de estudo em muitos campos, desde a psicologia à biologia, passando pela filosofia ou pela política, entre outras. Apesar deste facto não existe um conjunto acordado sobre quais as emoções básicas do ser humano. Em (Harré and Parrott, 1996) é afirmado que existem seis tipos de emoções primárias, nomeadamente amor, felicidade, surpresa, angústia, tristeza e medo, emoções estas que podem ser subdivididas em muitas emoções secundárias e terciárias, além de que cada emoção tem também diferentes intensidades. A força das opiniões está diretamente intricada com a intensidade de certas emoções. Contudo, opiniões e emoções não significam o mesmo.

Antes de analisar os sentimentos subjetivos expressos através de emoções e de opiniões, é fundamental realçar a distinção existente entre as duas noções seguintes: estado mental (ou estado de espírito) e as expressões linguísticas utilizadas para expressar este mesmo estado. Apesar do facto de existirem apenas seis tipos de emoções primárias, verifica-se, porém, a existência de um número significativamente maior de expressões linguísticas que podem ser utilizadas para expressar essas mesmas emoções. Além disso, existem, também, inúmeras expressões que denotam positividade ou negatividade numa expressão linguística, ou seja, a polaridade da expressão linguística em causa. Nesta situação, a análise de sentimentos ou a mineração de opiniões executa tentativas no sentido de inferir, e, consequentemente conhecer os sentimentos das pessoas, com base nas suas expressões linguísticas, determinando assim se o texto contém opiniões e sentimentos, positivos ou negativos, e até analisar outras partes do texto que possam constituir informação relevante para a aplicação prática destes resultados.

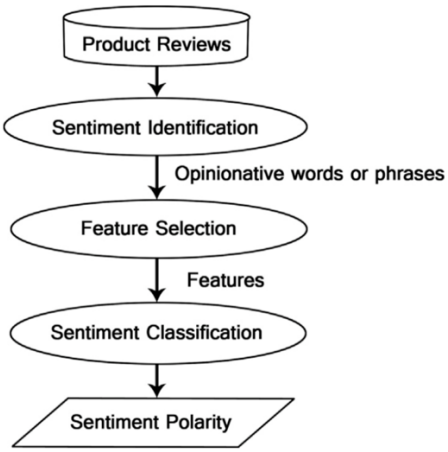


Figura 6: Processo de análise de sentimentos em críticas de produtos

Uma frase considerada objetiva traduz, de algum modo, uma informação factual sobre algo, enquanto que uma frase subjetiva expressa sentimentos, emoções ou pontos de vista pessoais. As expressões subjetivas tomam forma através de opiniões, alegações, desejos, crenças, suspeitas e especulações (Liu, 2010). Apesar disto, é importante enunciar, de igual modo, que nem todas as frases subjetivas contêm uma opinião. Tomemos como exemplo a frase: *"Eu pretendo um carro com um bom motor."* Esta frase não apresenta provas concretas da presença expressiva de uma opinião positiva ou negativa sobre qualquer carro ou sobre qualquer motor. De forma similar, note-se que uma frase objetiva também não exclui obrigatoriamente a presença de conteúdo opinativo. Exemplo disso mesmo é a frase *"A máquina de lavar avariou-se em menos de dois dias após a compra."*. Esta última frase é um claro exemplo de uma frase objetiva, mas que deixa, sem margem de dúvida, patente nela uma expressão de negatividade associada à prestação de um produto que num curto espaço de tempo, após a sua compra, deixou de funcionar corretamente. Sendo assim, podemos definir uma opinião explícita como uma opinião explicitamente contida numa frase subjetiva. Da mesma forma, podemos proceder à definição de uma opinião implícita como sendo uma opinião implicitamente expressa numa frase objetiva. Um exemplo de uma frase que explicita uma opinião positiva é *"A*

qualidade de imagem desta televisão é fantástica." E um exemplo de uma frase que expressa uma opinião negativa implícita é *"O mp3 avariou após três dias de uso."*. A última frase é claramente um facto objetivo, mas, apesar de tal, nela é expressa implicitamente uma opinião negativa sobre o *mp3*. Outro exemplo claro de uma frase objetiva que contém uma opinião negativa é "Após quatro dias de uso devolvi o mp3 à loja onde o comprei.". Esta frase demonstra explicitamente uma opinião negativa nela patente. O devolver do produto após meros quatro dias de uso serve de conclusão para a fraca qualidade do mesmo e para a subjacência do comentário que se encontra incluído na frase.

As frases objetivas contêm, assim, num nível subentendido, uma elevada carga de conteúdo opinativo, permitindo também derivar sentimentos das pessoas que, efetivamente, usufruem de um produto ou de um serviço e decidem partilhar a sua experiência com o mesmo. Por outro lado, define-se uma frase com conteúdo opinativo explícito como uma frase que de forma explícita ou implicitamente contém opiniões positivas ou negativas, podendo adquirir a forma de uma frase tanto objetiva como subjetiva.

Como demonstrado acima, os conceitos de frases subjetivas e de frases com conteúdo opinativo não são os mesmos. Não obstante este último facto, e por vezes, o conjunto das frases que contêm texto no qual está presente conteúdo opinativo constitui um subconjunto das frases subjetivas. A tarefa que se baseia em determinar se uma frase é subjetiva ou objetiva é a classificação de subjetividade. A noção de frase opinativa pode, indubitavelmente, ser aplicada a documentos (definidos como conjuntos de frases). No entanto, apenas se pode aplicar esta noção a documentos, se, de facto, os documentos possuem, pelo menos, uma frase com conteúdo opinativo. Exemplos desta definição são as críticas de produtos, e todo e qualquer *tweet* que expresse uma opinião sobre algo.

Cellular phone 1:
PHONE:
Positive: 125 <individual review sentences>
Negative: 7 <individual review sentences>
Feature: **voice quality**
Positive: 120 <individual review sentences>
Negative: 8 <individual review sentences>
Feature: **size**
Positive: 80 <individual review sentences>
Negative: 12 <individual review sentences>
...

Figura 7: Exemplo de sumário de opiniões baseadas em características de um produto.

Na Figura 7, explicitada em (Hu and Liu, 2004a), é possível observar que sobre um telemóvel exemplo foram analisadas várias críticas de produto, sendo que sobre o telemóvel em si, como um todo, contabilizaram-se cento e vinte e cinco frases que contêm uma opinião positiva e sete que contêm uma opinião negativa. Além das críticas sobre o produto em si, foram tidas em conta na avaliação duas características basilares do produto. Nessa avaliação, cento e vinte elementos avaliaram positivamente a qualidade de voz do produto e oito de forma negativa. Ainda quanto ao tamanho do telemóvel, oitenta críticas revelaram-se positivas e doze negativas. Na Figura 7 podemos ver uma possível apresentação desta primeira análise.

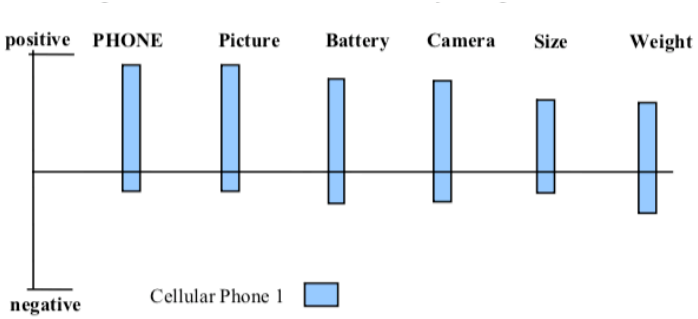


Figura 8: Visualização do sumário de opiniões baseadas em características do produto.

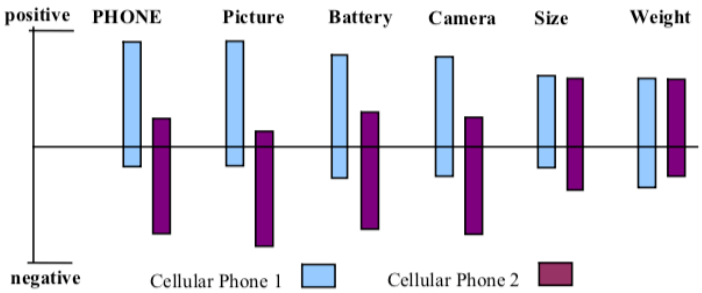


Figura 9: Comparação visual entre as opiniões sobre dois produtos.

Nas Figuras 8 e 9 (Hu and Liu, 2004a) podemos observar um sumário estruturado sobre as opiniões analisadas, na medida em que, graficamente, a retirada de ilações fica facilitada pela disposição e exposição dos dados da forma de apresentação utilizada. Na Figura 8 num espectro entre o positivo e o negativo é possível observar o valor obtido relativamente às diversas características do produto em análise. Esta visão apresentada permite saber quais as características do produto que obtiveram melhor aceitação por parte do público alvo. Na Figura 9, e mantendo o espectro entre o positivo e o negativo, procede-se à comparação entre dois produtos da mesma categoria, neste caso em específico, telemóveis, procurando fatores comuns aos dois, ou seja, partes constituintes dos dois passíveis de serem avaliadas. Esta comparação visual permite-nos avaliar as melhores e as piores características (Zhang and Zhu, 2013) de um em relação ao outro. É uma comparação interessante, tanto num cenário em que ambos os telemóveis são produzidos pela mesma empresa, ou por uma empresa diferente. Neste caso, também são providenciados os cenários necessários para a comparação entre mais telemóveis, permitindo avaliar o que melhor se adequa a um utilizador em particular (Najmi *et al.*, 2015). Além disso, também releva especialmente a perspetiva empresarial de quem vende o produto e na perspetiva individual do cliente que efetua ou pretende efetuar a compra de um ou dos dois dispositivos. A comparação entre vários produtos de um mesmo

segmento de mercado revela-se de especial interesse, na medida em que, com isso, se apuram as qualidades e defeitos de uma perspetiva comparativa. Adicionalmente, traz grandes benefícios, quer em termos empresariais, para as empresas que atuam nesse mesmo mercado, quer para os clientes, que pensam em adquirir produtos como esses.

A título de exemplo, na Figura 9 pode observar-se que, no papel de um cliente interessado, o telemóvel 1 destaca-se do telemóvel 2, pelos comentários positivos em relação à qualidade de imagem, bateria e qualidade da câmara. Este tipo de avaliação é cada vez mais, nos dias de hoje, preponderante nas tomadas de decisão dos possíveis clientes. Em vários mercados é possível retirar o maior proveito das *críticas* feitas pelos clientes desse mesmo mercado e sobre as empresas que nele atuam. Em (Wu *et al.*, 2013) foi desenvolvida uma visualização interativa das *críticas* feitas pelos clientes de um hotel. Desse modo o hotel conseguiu ter o *feedback* necessário para colmatar possíveis lacunas existentes nos seus serviços ou procedimentos. Com coisas como essa, é possível distinguir um sumário a nível de características do produto ou de todo o produto em si, o que permite a construção e implementação de um *timeline* e consequente *tracking* das tendências e evoluções de todo e qualquer mercado, servindo os interesses de quem produz e de quem adquire os produtos ou serviços.

3.5 Classificação de Sentimentos

A polaridade dos dados implica, consequentemente, a classificação dos sentimentos encontrados como resultado da análise de texto executada. A classificação de sentimentos e subjetividade é um tópico bastante estudado. Esta

baseia-se na classificação de texto que contém opiniões, como são o caso das *críticas* de produtos que expressam, na maior parte das vezes, opiniões negativas ou positivas (Das and Chen, 2007). Esta tarefa é, de igual modo, conhecida como classificação de sentimentos ao nível dos documentos, uma vez que considera todo um texto ou um documento como unidade básica de informação.

A mesma classificação de sentimentos pode ser aplicada naturalmente a frases individuais. Todavia, a tarefa que consiste na classificação de uma frase como contendo uma opinião ou não, denomina-se por classificação de subjetividade. Por outro lado, as frases que apresentam carácter opinativo podem ser posteriormente submetidas a uma análise de polaridade, ou seja, uma análise que determina se as opiniões presentes no texto se baseiam em expressões positivas ou negativas. Esta análise é denominada por classificação de sentimentos ao nível frásico (Liu, 2010).

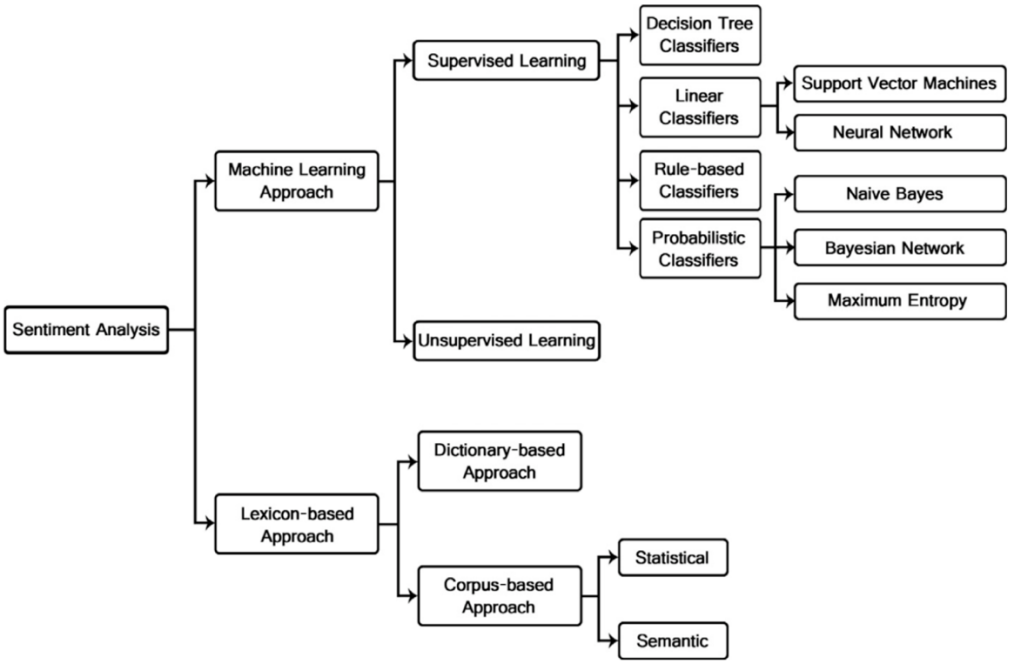


Figura 10: Técnicas de classificação de sentimentos.

A maior parte das técnicas usadas na classificação de sentimentos ao nível de documentos baseia-se em aprendizagem supervisionada, embora existam também métodos de aprendizagem não supervisionada, tal como se pode verificar pela análise da Figura 10. Além do interesse académico em estudar sistemas de classificação de sentimentos eficientes, também existe interesse nestes mesmos sistemas por parte das aplicações de Business Intelligence (Terveen *et al.*, 1997)(Tatemura, 2000), nas quais o input recebido por parte dos utilizadores pode ser eficientemente sumariado.

3.5.1 Classificação de Sentimentos Baseada em Aprendizagem supervisionada

A aprendizagem supervisionada é a aprendizagem realizada a partir de exemplos, na maior parte das vezes orientada por um analista de dados em que o papel deste consiste no auxílio na construção do modelo preditivo, através da definição de classes e dos exemplos que tem como resultado determinada classe. O sistema a ser treinado determina a descrição para cada classe, sendo esta o conjunto de características similares de cada conjunto, pertencente a cada classe, que vão sendo expostos nos exemplos que lhe vão sendo fornecidos. Após a determinação da descrição, torna-se possível a formulação da regra de classificação que pode ser usada para predizer a classe de elementos que respondem aos casos do conjunto de elementos pertencentes ao grupo de teste. A classificação de sentimentos pode claramente ser tida e formulada como um problema ao qual a aprendizagem supervisionada pode responder, definindo desde início duas classes distintas, nomeadamente a positiva e a negativa. O

conjunto de dados de teste pode, usualmente, ser um conjunto de críticas sobre produtos.

Um exemplo de classificação de texto noutra vertente e usada para diferentes finalidades é a classificação de texto por tópicos, ou seja, classificar documentos de acordo com a sua pertença a classes como sejam o desporto, a sociedade, a cultura, a música ou a arte, entre outras. Contudo, na classificação de texto, que possui como objetivo a análise de sentimentos as palavras que indiciam a relação do texto com tópicos como os anteriormente enumerados não são as que têm uma relevância principal. Porém não deixam, como é óbvio, de ser importantes. Ao invés disso, as palavras implicitamente relacionadas com a expressão de opiniões e de sentimentos são maioritariamente analisadas, uma vez que estas indicam a presença de opiniões negativas ou positivas (Chesley *et al.*, 2006). Exemplo de palavras como essas, num primeiro nível são: fantástico, mau, horrível, ou estupendo, entre muitas outras.

Vários métodos de aprendizagem supervisionada podem ser aplicados no sentido de executar a classificação de sentimentos. De referiri, por exemplo, o *Naive Bayes* e os *SVMs (Support Vector Machines)*. Em (Metsis, Androutsopoulos and Paliouras, 2006) é analisada a capacidade de filtragem de *emails* que contêm spam, utilizando *Naive Bayes* enquanto que em (Pang, Lee and Vaithyanathan, 2002) foram utilizadas técnicas de aprendizagem supervisionada foram utilizadas para classificar críticas de filmes em duas classes, positiva e negativa. Recorrendo ao uso de palavras individuais, como sendo características do tipo de frases no qual se pretendia mapear cada frase analisada, a classificação foi realizada com sucesso tanto usando *Naive Bayes* como *SVM*. As críticas neutras não foram tidas em conta no estudo, tornando menor a complexidade do problema em questão.

Tal como a maior parte das aplicações de *Machine Learning*, a tarefa basilar reside em arquitetar um conjunto de características ajustadas à resolução do problema. Em (Pang and Lee, 2008) são enunciados alguns desses conjuntos tais como os termos e a sua frequência, as etiquetas de partes do discurso, as palavras opinativas em frases, a dependência sintática, e a negação. Os termos

e a sua frequência são características que tomam forma em palavras individuais ou *n-grams*[8] e a sua respetiva contagem. As etiquetas de partes do discurso são partes do discurso que denotam especial expressividade característica do indivíduo que o formula. Os adjetivos são indicadores importantes da presença de subjetividade e consequentemente de opinião/texto opinativo. As palavras de opinião são palavras que, usualmente, são utilizadas para expressar sentimentos positivos ou negativos. Exemplificando, as palavras bonita, estupendo, bom, e fantástico são palavras que auxiliam na construção de uma opinião positiva. De forma contrária, as palavras mau, paupérrimo, e terrível são palavras que caracterizam uma opinião negativa. Não obstante o facto de que a maioria das palavras opinativas são adjetivos (e.g. lindo, feio, mau, bom, fantástico ou terrível), os advérbios (e.g. jamais, nunca, realmente, positivamente ou negativamente), os nomes (e.g. lixo, jóia ou obra) e os verbos (e.g. gostar, desgostar ou odiar) possuem a plena capacidade de expressar opiniões. Por outro lado, também são de grande importância as frases características de idiomas em específico tais como *"O carro custou-me os olhos da cara."* ou *"Nem que me pagassem assistiria a esse filme"*.

A dependência sintática refere-se ao facto de as frases possuírem uma organização interna, de tal forma que o estabelecimento de árvores de dependência se revela possível e de grande utilidade na análise da frase. A negação destaca um estágio da análise que é, de modo similar, importante, uma vez que é nesta etapa que a polaridade de uma frase pode, de modo significativo sofrer alterações. A presença de uma palavra que denota negação nem sempre significa uma opinião negativa, podendo mesmo até não significar um peso negativo na polaridade da frase analisada. As palavras que denotam negação são fundamentais uma vez que a sua aparência e a forma como são aplicadas podem alterar a orientação da frase em análise, como acontece, por exemplo, nas frases

[8] *N-grams* é uma sequência contígua de *n* itens retirados a partir de um texto. Os itens podem adquirir a forma de fonemas, sílabas, letras, ou palavras.

"Nunca vi um telemóvel tão bom." ou *"Não só gostei da câmara do telemóvel como também gostei do facto de possuir 3D touch."*.

Outro fator importante aquando da realização da classificação de sentimentos usando aprendizagem supervisionada é o facto de treinar a classificação de sentimentos tendo em conta um certo domínio e aplicar dados de teste que pertencem a outro domínio que pode influenciar da pior forma os resultados obtidos. A razão para tal prende-se com o facto de que as palavras, e até a linguagem utilizada para construir opiniões, poderem variar de domínio para domínio. Pode mesmo até acontecer que a palavra utilizada para denotar positividade num domínio, esteja por sua vez associada a sentimentos de negatividade noutro domínio completamente diferente. Observe-se, a título de exemplo o desempenho e a expressividade da palavra *"imprevisível"* em duas críticas de domínios diferentes a seguir apresentadas.

- *"O filme tem um guião imprevisível fazendo com que o telespectador fique bastante cativado."*
- *"O telemóvel tem um desempenho de bateria imprevisível."*

Na primeira frase a palavra *"imprevisível"* possui uma polaridade positiva, na medida em que define um grande valor de interesse no que toca à qualidade do filme e ao grande interesse que este desperta no telespectador. Já na segunda frase a palavra "imprevisível" destaca-se exatamente pelo seu sentido literal, a imprevisibilidade, mas no sentido negativo, tendo em conta o contexto em que é aplicada. Na frase caracteriza-se a bateria do telemóvel que se encontra em avaliação pela sua imprevisibilidade, o que é um ponto claramente negativo relativamente à avaliação de uma bateria de um telemóvel, que, em casos normais, se espera que seja bastante regular e que permita providenciar ao telemóvel uma autonomia considerável de funcionamento. Em resumo, a adaptação do modelo a um domínio em específico é um requisito a ter na

construção do mesmo, para que se possam esperar bons resultados da aplicação deste a dados de teste.

3.5.2 Classificação de Sentimentos Baseada em Aprendizagem não supervisionada

Os indicadores dominantes na classificação de sentimentos são claramente todo o conjunto de palavras opinativas e as frases em que estas se inserem e que são utilizadas com poderes expressivos. A aprendizagem não supervisionada baseia-se nestes fatores para a construção dos modelos que resultam da sua aplicação. Em (Peter D. Turney, 2001) está apresentado um algoritmo de aprendizagem não supervisionada para classificação de críticas do estilo *"thumbs up"* e *"thumbs down"*. Inicialmente é pressuposto que a classificação de cada *review* é calculada como a média da orientação semântica das frases que a *review* em si contém, dando especial enfoque aos adjetivos e advérbios. No algoritmo o cálculo da orientação semântica da frase é realizado como a informação mútua[9], entre a frase recebida como *input* e a palavra excelente, menos a informação mútua entre essa mesma frase e a palavra "pobre" (Church and Hanks, 1989). O algoritmo referido alcançou uma eficácia de 74% (setenta e quatro por cento) aquando da avaliação de 410 (quatrocentas e dez) *críticas* retiradas do *Epinions*[10], e abrangendo quatro domínio diferentes, nomeadamente avaliações de automóveis, de bancos, de filmes e de destinos de viagem. O espectro do sucesso de desempenho e eficácia do algoritmo localizou-se entre os

[9] De acordo com *(Fano, 1961)* (Fano and Hawkins, 1961), se dois pontos (palavras), *x* e *y*, possuem probabilidades *P(x)* e *P(y)* respetivamente, a sua informação mútua denotada por *I(x,y)* é definida como $I(x,y) \equiv log_2 \frac{P(x,y)}{P(x)P(y)}$.

[10] O site epinions.com foi um site de críticas de produtos, tendo sido adquirido pelo shopping.com que por sua foi comprado pelo Ebay em 2005.

84% (oitenta e quatro por cento), no domínio automóvel e os 66% (sessenta e seis por cento), no domínio filmográfico. A primeira fase deste algoritmo contempla a extração de frases que contém adjetivos ou advérbios, uma vez que estes elementos se encontram entre os melhores indicadores da polaridade de um texto e da carga expressiva que este carrega, traduzindo as opiniões e a subjetividade que nele estão contidos. Apesar deste facto, um adjetivo quando isolado não é suficiente para a determinação da polaridade do texto de uma forma correta, uma vez que, não se encontrando num cenário bem contextualizado a sua interpretação relativamente à carga de polaridade que expressa não pode ser determinada com certeza. O algoritmo em questão extrai duas palavras consecutivas, nas quais uma delas é o adjetivo/advérbio e a outra é a palavra que o contextualiza no texto em questão. Vejamos de uma forma simples como o processo se desenrola. Primeiramente um etiquetador *POS* é aplicado ao texto constituinte da crítica. Depois, procede-se à extração das duas palavras, se as suas etiquetas *POS (Part of Speech)* se mapeiam em algum dos padrões apresentados na Figura 11 - a etiqueta *JJ* corresponde aos adjetivos, a *NN* corresponde aos nomes, a *RB* aos advérbios, e a *VB* aos verbos (Santorini, 1990).

	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

Figura 11: Padrões de etiquetagem para extração de duas palavras das críticas.

Deste modo, o algoritmo extrai duas palavras consecutivas, quando um membro do par é um adjetivo ou um advérbio e a outra metade quando um membro é uma palavra relativa ao contexto. Duas palavras consecutivas são extraídas se as suas etiquetas *POS* correspondem a um padrão explicitado na Figura 11. A título de exemplo, note-se o padrão apresentado na linha 1: duas palavras consecutivas são extraídas, se a primeira palavra é um adjetivo e a segunda palavra é um nome, sendo que a terceira palavra (que não é extraída) pode ser qualquer elemento constituinte da frase.

Na frase exemplo *"Este carro proporciona uma ótima condução."*, as palavras *"ótima condução"* são extraídas como satisfazendo a condição 1, uma vez que a primeira palavra é *"ótima"* – um adjetivo -, e a segunda palavra é *"condução"* – um nome. Na segunda etapa do algoritmo, é feita a estimativa da polaridade dos dados contidos nas frases extraídas, usando a informação mútua como forma de avaliar e quantificar como uma medida real a força ou a intensidade da associação semântica entre duas palavras (Church and Hanks, 1989). *(Church & Hanks, 1989)* propuseram esta medida alternativa como sendo um rácio associativo, como uma forma de medir as normas de associação entre palavras, baseando-se no conceito teorético da informação denominado de informação mútua. Além disso, afirmaram ainda a maior objetividade desta medida e o reduzido custo a ela associado quando comparada com o método subjetivo apresentado por (Wicklund, Palermo and Jenkins, 1964).

O conceito de informação mútua (Fano and Hawkins, 1961) é definido através da definição de dois pontos (palavras), *x* e *y*, que têm como probabilidades *P(x)* e *P(y)* respetivamente, sendo a sua informação mútua denotada por *I(x,y)* e definida da seguinte forma:

$$I(x,y) \equiv log_2 \frac{P(x,y)}{P(x)P(y)} \qquad \text{(Fórmula 1)}$$

A informação mútua compara a probabilidade de observar *x* e *y* juntos (probabilidade conjunta) com as probabilidades de observar *x* e *y*

independentemente (hipótese). Se, de facto, existir uma associação genuína entre x e y, a sua probabilidade conjunta *P(x,y)* será muito maior do que *P(x)P(y)*, e consequentemente I*(x,y)* $\gg$ *0*. Se não existe nenhuma relação significativa entre *x* e *y*, então *P(x,y)* $\approx$ *P(x)P(y)*, e daqui, *I(x,y)* $\approx$ *0*. Por fim, se *x* e *y* se encontram numa distribuição complementar, então *P(x,y)* será muito menos do que *P(x)P(y)*, e então *I(x,y)* $\ll$ *0*.

O algoritmo *PMI-IR* foi empiricamente avaliado aquando da sua aplicação sobre 80 questões do teste de sinónimos do *TOEFL*[11], obtendo um resultado relativamente à sua eficácia de 74% (Peter D Turney, 2001). Para fins de termo de comparação foi utilizado o algoritmo *LSA* (Análise Semântica Latent) (Landauer, Folt and Laham, 1998), uma outra forma de medida estatística para a associação de palavras, que atingiu um resultado de 64% tendo como base as mesmas questões (Landauer and Dumais, 1997) aplicadas ao algoritmo *PMI-IR*.

A *Pointwise Mutual Information (PMI)* entre duas palavras, palavra$_1$ e palavra$_2$ é definida da seguinte forma (Church and Hanks, 1989):

$$PMI(palavra_1, palavra_2) = log_2 \left[\frac{p(palavra_1 \wedge palavra_2)}{p(palavra_1)p(palavra_2)}\right] \quad \text{(Fórmula 2)}$$

Na Fórmula (2), *p(palavra$_1$* $\wedge$ *palavra$_2$)* denota a probabilidade da *palavra$_1$* e da *palavra$_2$* co-ocorrerem. Se as palavras são estatisticamente independentes então a probabilidade da sua co-ocorrência é dada pelo produto entre *p(palavra$_1$)* e *p(palavra$_2$)*. O rácio entre *p(palavra$_1$* $\wedge$ *palavra$_2$)* e a *p(palavra$_1$)p(palavra$_2$)* é a medida do grau de dependência estatística entre as palavras. O *logaritmo* deste mesmo rácio é a quantidade de informação que se adquire sobre a presença de uma das palavras quando se observa a outra.

[11] TOEFL – Test of English as a Foreign Language, traduz-se como teste do Inglês como Língua Estrangeira.

Por sua vez, a orientação semântica/polaridade da opinião *(Semantic Orientation - SO)* (Peter D. Turney, 2001) de uma *frase* é calculada através da Fórmula (3):

SO(frase) = PMI(frase, "excelente") – PMI(frase, "pobre") *(Fórmula 3)*

As palavras de referência *"pobre"* e *"excelente"* foram escolhidas devido à regular escolha destas para a definição de uma crítica de uma estrela e de cinco estrelas, respetivamente. No seguimento da definição da *Fórmula (3)* conclui-se que a orientação semântica é positiva quando a frase é mais fortemente associada com a palavra *"excelente"*, e negativa quando a frase é mais fortemente associada com a palavra *"pobre"*.

Na terceira e última fase do algoritmo, este recebe uma crítica como input, calcula a orientação semântica média, tendo em conta todas as frases que constituem a crítica que recebeu como input, e classifica a crítica como recomendada, se a média da orientação semântica for positiva, ou como não recomendada no caso contrário.

Em (Dave *et al.*, 2003) foi proposta uma ferramenta de mineração de opiniões que, idealmente, processa um conjunto de resultados de pesquisa para um dado item, gerando uma lista de atributos do produto (qualidade, características, entre outros) e agrega as opiniões existentes sobre cada um deles (pobre, neutro ou bom). Com esta ferramenta, primeiramente, realiza-se a identificação de propriedades únicas referentes a este problema e é desenvolvido um método para, automaticamente, distinguir as críticas positivas das negativas. O classificador começa por realizar, através de técnicas de pesquisa de informação, a extração de características, atribuindo-lhes um resultado, e implementando assim resultados para as várias métricas, dependendo das heurísticas da situação de teste. É referido também o ruído e a ambiguidade provocado por extrações realizadas em motores de pesquisa *Web* como um

obstáculo para a boa performance do modelo. Apesar de tal, no contexto do estudo verificou-se a obtenção de resultado qualitativamente úteis.

Assim, é notório que a aprendizagem não supervisionada revela-se de enorme utilidade na mineração de opiniões e sentimentos, sendo que, para uma execução com sucesso de um sistema de mineração de opiniões e sentimentos, especificamente neste contexto, requer uma preparação específica para que o modelo esteja preparado e implementado de modo a adequar-se ao cenário onde o trabalho se encontra a ser realizado.

3.6 Índices de Satisfação

Belo e Carvalho (2016) apresentaram para um sistema de gestão de filas uma forma eficiente de o fazerem, que consiste em ter sempre filas com reduzido comprimento e um serviço de atendimento rápido. Todavia, este é um objetivo difícil de atingir devido a vários fatores. Além disso, com o passar do tempo, e na maior parte dos casos, os serviços de atendimento têm tendência a deteriorar-se, especialmente se o número de pessoas para serem atendidas aumentar, ficando assim o nível de eficácia de atendimento reduzido, tal como é referido em (Belo e Carvalho, 2016). Neste estudo foram analisadas e implementadas várias abordagens para prever o tempo estimado de espera, dada uma linha cronológica relativamente à história do serviço de atendimento, e como este poderia evoluir no futuro. Para tal foi usado um conjunto de dados reais extraídos, através de uma companhia de telecomunicações específica, que possuía problemas reais existentes no serviço de atendimentos aos seus clientes. A partir daqui, podemos observar este caso real como um cenário no qual os índices de satisfação poderão ser de grande proveito, uma vez que permitem

estimar com maior assertividade o grau de satisfação obtido por parte da empresa juntos dos seus clientes, relativamente aos processos inerentes em avaliação, nomeadamente o apoio ao cliente.

Basicamente, um índice de satisfação é um indicador que serve de orientação, principalmente para empresas, mas também para consumidores, ou seja o alvo que as empresas objetivam como o possível consumidor dos seus produtos ou serviços. Um índice de satisfação permite, através de uma perspetiva intuitiva, o fornecimento de níveis de satisfação que podem ser compreendidos como métricas de avaliação de um negócio, e que podem ser analisados sob diferentes fatores de avaliação, para que se compreenda a opinião que, por exemplo, o mercado tem sobre um produto que acabou de ser lançado por uma empresa. Os níveis obtidos, relativamente aos fatores de avaliação existentes, determinam o espaço que existe para melhorar e o que deve ser feito para melhorar, nomeadamente em casos nos quais a empresa procura obter junto dos seus clientes, opiniões sobre os produtos que comercializa. Por exemplo, no caso de um espaço de suporte ao cliente pode ser instituída a existência de fatores avaliativos, como a rapidez do atendimento, a eficácia do atendimento ou o nível de satisfação após atendimento, e, no caso de se pretender avaliar a recetividade do mercado ao lançamento de um telemóvel, pode recorrer-se a questões feitas a clientes que tenham experimentado o telemóvel, nomeadamente sobre a usabilidade do mesmo, sobre o modo de utilização do telemóvel e quão fácil é executar determinadas tarefas inerentes à utilização do telemóvel, como por exemplo ligar a alguém, ou alterar a definição de brilho do ecrã. Além disso pode-se igualmente questionar-se sobre a autonomia do aparelho ou sobre a durabilidade do aparelho, entre muitos outros fatores.

A resposta a estas questões colocadas a uma amostra de clientes vai permitir a implementação e obtenção de índices de satisfação. Se as respostas forem classificadas em bom, médio, ou mau, e tendo em conta o exemplo acima descrito sobre a avaliação de um telemóvel, pode-se, a título de exemplo, obter um índice de satisfação relativamente à autonomia obtida pelo telemóvel. Este

mesmo índice de satisfação descreve um cenário que permite avaliar se terão de ser realizadas bastantes melhorias no sentido de aumentar a autonomia do telemóvel ou se a autonomia do produto foi bem conseguida. Com isso, poderemos obter, por exemplo, um gráfico no qual 70% dos inquiridos qualificou a autonomia como boa, 20% como tendo um desempenho médio, e 10% como tendo um desempenho mau. Tal disponibilização de resultados permite a avaliação do cenário através de duas perspetivas: a da empresa fornecedora do produto ao mercado, e a do cliente ou analista de mercado. Por parte da empresa fornecedora do produto ao mercado, tais resultados permitem avaliar fielmente a qualidade do trabalho que está a ser desenvolvido e quais os fatores em que têm de apostar mais para melhorar cada vez mais os seus telemóveis. Por parte dos clientes, pode-se ter esta disponibilização de resultados como uma vantagem, por exemplo, para possíveis novos clientes. Um novo cliente ao consultar estes resultados verifica que, em relação à autonomia deste telemóvel, a maior parte dos inquiridos no estudo classifica-a como boa, e, portanto, este novo cliente vai certamente incluir no lote das suas possíveis escolhas para compra de um telemóvel este mesmo produto analisado, aumentado assim o lucro da empresa, se o processo assim se realizar repetidamente.

A influência que estas análises possuem no mercado são significativas, uma vez que, atualmente, e em virtude de muitos mercados possuírem grande oferta, tal como é o caso do mercado dos telemóveis, grande parte das pessoas recorre à análise de resultados tais como este acima mencionado para escolher o seu próximo telemóvel. De igual modo, cada telemóvel é analisado especificamente em relação a vários fatores e através de questões colocadas aos consumidores como modo de calcular o índice de satisfação destes sobre o produto (Figura 12), permitindo analisá-lo através de diferentes perspetivas e, assim prever o seu desempenho.

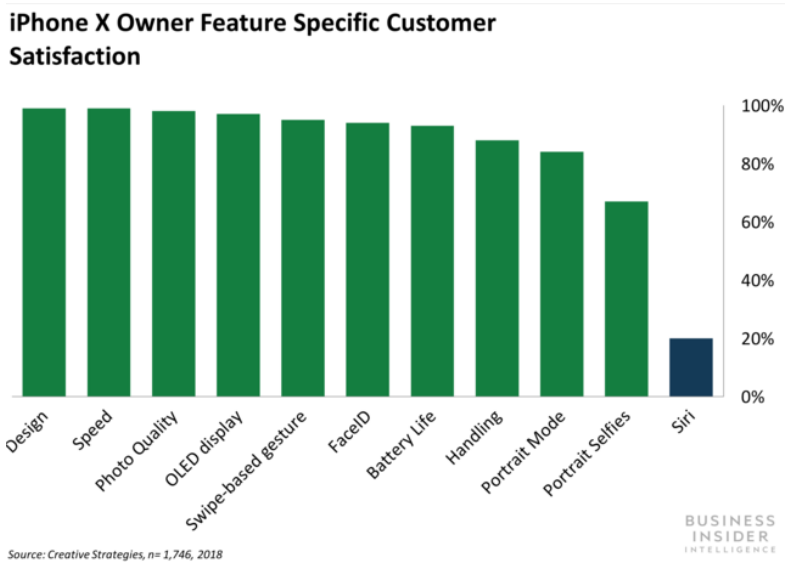


Figura 12: Análise de vários fatores sobre o iPhone X (Fonte: Creative Strategies, n=1746, 2018).

A análise revelada na Figura 12 permite a novos clientes analisarem se a compra deste telemóvel é viável ou não, ao mesmo tempo que providencia fielmente à empresa que o produz, as opiniões dos consumidores sobre este produto, o que permite analisar que fatores necessitam de melhorias e que fatores mais satisfazem os clientes. Os dados apresentados na Figura 12 providenciam, em contexto empresarial, uma fundamentação significativa e bem estruturada para as tomadas de decisão na produção do produto que se segue a este que se encontra em análise. No próximo produto, muito provavelmente e teoricamente, a empresa tenta melhorar os aspetos menos bons deste atual, e continua a desenvolver o bom trabalho relativamente aos pontos melhor tidos em consideração pelos consumidores.

Em (Fornell *et al.*, 1996) é possível verificar o porquê do aparecimento do ACSI (Índice Americano de Satisfação do Consumidor), bem como o proveito que tal apresenta para a condução de estudos de performance de mercado através dos dados registados ao longo do tempo.



Figura 13: Índice de Satisfação sobre o tema brexit retirado da aplicação OpClassify implementada e apresentada no Capítulo 4.

Na aplicação implementada (Capítulo 4) será possível, de igual modo, verificar a utilidade de vários resultados representativos de índices de satisfação. Um exemplo de índice de satisfação é possível ser observado na Figura 13, sobre o tema *"brexit"* e tendo em conta as opiniões partilhadas na plataforma *Twitter* que são analisadas pela aplicação OpClassify apresentada no Capítulo 4.

4 Um Caso de Estudo

O caso de estudo que apresentamos neste capítulo envolve a implementação de uma aplicação de extração de dados e a sua posterior análise de dados. Com este caso pretende-se demonstrar a aplicação da mineração de opiniões e de sentimentos em prol da indústria ou do mercado, quer seja para empresas que atuam no mercado ou para os próprios clientes-consumidores dessas mesmas empresas. Na implementação deste caso demos grande relevo no fornecimento de dados analisados, tendo estes o intuito de servir de fundamento para a tomada de decisões a nível de qualquer agente de mercado. Nas secções seguintes especificar-se-á pormenorizadamente todos os estágios do desenvolvimento da aplicação em causa, de modo a explanar e a providenciar, do melhor modo, a forma e os procedimentos que regeram a sua implementação da aplicação em questão.

4.1 Contextualização da Aplicação

Como é sabido, o processo de análise de sentimentos é tecnicamente desafiante, mas também repleto de enorme utilidade prática. Em qualquer empresa que atue num mercado de qualquer tipo, é usual verificarmos que existe uma vontade de saber o que o público-alvo perspetiva, em relação ao uso de um serviço ou de um qualquer produto adquirido através dessa mesma empresa. Uma empresa competitiva e que procura o contínuo progresso deve tomar cada cliente como um dado importante em qualquer processo de análise sobre os seus clientes. Uma empresa deve utilizar esses dados de modo a aperfeiçoar a sua

atividade, a sua produção, os seus modos de desenvolvimento dos seus produtos, e tudo aquilo que possa ser aprimorado a partir do *feedback* recebido a partir dos clientes.

O estudo computacional das opiniões, avaliações, atitudes e emoções das pessoas, em relação a entidades, indivíduos, questões, eventos, tópicos e os seus respetivos atributos (Liu and Zhang, 2012), acrescentam um valor imensamente significativo aos dados que são recolhidos pelas empresas. Usar estes dados para fazer crescer ou desenvolver uma empresa é um modo sustentado e bem fundamentado para induzir o progresso empresarial no mercado. A grande competitividade existente no mercado aumenta o interesse em conhecer os clientes de forma a poder corresponder aos seus gostos de forma cada vez melhor, tentando com isso diminuir o número de pessoas com opinião negativa ou insatisfeita com os produtos ou com os serviços da empresa. Além do mais, e de modo cada vez mais frequente, é por parte dos consumidores que parte a vontade de conhecer e obter o máximo de informação no que respeita às opiniões de outros clientes que adquiriram o produto que estes novos clientes pretendem agora adquirir. Liu (2010) categorizou a informação textual existente em dois grupos basilares: os factos e as opiniões. O conjunto dos factos inclui as expressões objetivas sobre entidades, eventos e suas propriedades, enquanto que o conjunto das opiniões compreende, regularmente, e na sua maior parte, expressões subjetivas que descrevem os sentimentos das pessoas, as avaliações ou os sentimentos perante entidades, eventos e as suas propriedades.

Antes do aparecimento da *(Web)* quando um indivíduo pretendia tomar uma decisão sobre a compra de um produto ou obtenção de um serviço, perguntava a conhecidos ou amigos a sua opinião sobre esse produto ou serviço. No caso de um grupo empresarial, quando este pretendia adquirir uma nova perspetiva sobre a opinião que os seus clientes possuíam sobre os seus produtos que este comercializava, a empresa procurava realizar inquéritos e outro tipo de pesquisas. Atualmente, a forma como as pessoas declaram as suas opiniões e os seus sentimentos mudou. Hoje é mais simplificada, tomando frequentemente a

forma de uma crítica num *website* de críticas de restaurantes ou de um simples *tweet*, entre outros. Qualquer pessoa partilha críticas dos produtos que adquiriu e publica os seus pontos de vista sobre quase tudo em *blogs*, em grupos de discussão ou em fóruns. A nova forma das pessoas expressarem descontentamento, aborrecimento, alegria, felicidade e qualquer outro estado de espírito traduz um novo comportamento que permite definir um conjunto de novas métricas sobre a informação, e, sobretudo, provoca o aparecimento de novas fontes de informação com várias aplicações práticas.

Nos dias de hoje, quando se pretende adquirir um produto, e, antes da aquisição se pretender conhecer melhor esse produto, vai-se, muito provavelmente, consultar-se as críticas existentes na *Web* sobre esse mesmo produto. Numa perspetiva empresarial, com toda esta revolução de comportamento social e o modo como se passam a expressar sentimentos e opiniões, já não é mais necessário despender dos próprios recursos para conhecer a opinião dos seus clientes sobre os seus produtos ou serviços, ou para saber as opiniões do mercado sobre os produtos de qualquer empresa sua concorrente. A existência de inúmeros *blogs* de *críticas*, fóruns de discussão, e redes sociais (*Facebook, Instagram, Twitter, etc.*), transformaram as opiniões públicas expressas em todo o tipo de meios acima e providenciaram um aumento do efeito e da influência destas na sociedade em geral. A quantidade de dados gerada por toda esta atividade de expressão opinativa, exercida no contexto de um mercado é bastante elevada e tem grande valor empresarial. Assim, facilmente se nota que as tarefas que incluem a pesquisa, organização, monitorização e interpretação desses dados, para que se consigam efeitos benéficos junto dos objetivos empresariais, são operações muito complexas e trabalhosas. Tal deve-se ao facto de que, além da tarefa de encontrar estes mesmos dados, adiciona-se a tarefa de pré-processar, processar, e, mais especificamente, interpretar os dados obtidos. A adicionar a toda esta complexidade, note-se, de igual modo, a complexidade do próprio processo de análise de tais dados.

Para cada sítio *Web*, existem enormes quantidades de texto que possuem características e perspetivas, na sua maioria, extremamente específicas relativamente ao contexto em que estão inseridas. Desta forma, e na grande maioria dos casos, estes sítios hospedam opiniões bastante intricadas e complexas no que à sua interpretação diz respeito, tendo em vista a sua categorização e consequente determinação da sua utilidade.

Consequentemente, após todo este processo, o conhecimento específico que se retira da pesquisa e análise das *críticas* pode não ser o melhor possível, podendo mesmo não representar utilidade. Além do mais, com um volume tão grande de informação, a extração de conhecimento útil dessa mesma informação torna-se uma tarefa deveras complexa. Tendo isto em conta, a aplicação tenta, desde modo, implementar um processo de mineração de opiniões e de sentimentos, com a adição de um sistema de sumarização, uma vez que, assim, se poderão retirar enormes benefícios a múltiplos níveis, tanto para clientes como para as empresas em termos gerais. A perspetiva providenciada relativamente aos temas que a aplicação recebe pretende, assim, servir de base para um processo de tomada de decisões ajustado e eficiente.

4.2 A Motivação para a Implementação da Aplicação

Medhat, Hassan and Korashy (2014) definiram a análise de sentimentos ou a mineração de opiniões como sendo o estudo e a consequente tentativa de interpretação computacional das opiniões, atitudes e emoções das pessoas. Este tipo de processo referente à interpretação computacional das opiniões de cada pessoa, permite a existência de sistemas que fazem das opiniões e sentimentos partilhados na *Web* matéria-prima bastante preciosa*,* para, deste modo, criar e

traçar o gráfico que evidencia o bom recebimento de um tema na sociedade ou o seu exato contrário.

Todos estes elementos são dados que permitem a criação de bases sólidas para melhorar os serviços prestados aos clientes, numa perspetiva empresarial, e para melhorar os estudos de mercado tanto efetuados por empresas ou por agentes de mercado individuais. A nível empresarial, os dados recolhidos junto das situações em que as empresas interagem com os clientes, como por exemplo em situações de prestação de suporte, são muito preciosos, na medida em que manifestam uma imensa utilidade para suportar melhorias no serviço a longo e a curto prazo, consoante a ilação retirada a partir da análise dos dados disponibilizados. Assim, o nível de performance de uma empresa usualmente tende a melhorar, conservando-a num bom nível de satisfação dos seus clientes, cativando possivelmente novos consumidores.

A nível individual existem vários exemplos através dos quais a análise dos dados facilita a tomada de várias decisões que surgem a qualquer pessoa que esteja a analisar várias hipóteses para adquirir determinado produto e/ou serviço. Na maior parte dos casos, quando alguém quer comprar um computador costuma analisar os pontos mais característicos de um computador e a partir daí procurar a marca que melhor conjugue a maior parte desses pontos num produto só, ao mesmo tempo que estima um nível preço/qualidade razoável. Várias são as ofertas que o mercado nos faz. Por isso mesmo, uma análise dos dados mais intuitiva e cuidada facilita a tomada de decisão por parte do utilizador de modo a que nenhum aspeto da compra seja descurado, permitindo avaliar além do mais, as críticas realizadas por consumidores que já experimentaram o produto e que podem classificar de melhor forma os seus pontos positivos e negativos.

Em (Devitt and Ahmad, 2007) a identificação da polaridade dos sentimentos existentes nas opiniões partilhadas em sítios *Web* provou ser de grande influência na venda ou compra de ações de uma dada empresa, tendo em conta a melhor cotação sentimental obtida nas várias críticas existentes e analisadas pelos interessados no mercado. Desta forma, e recorrendo à

mineração de opiniões e sentimentos, um negócio pode, indubitavelmente, suportar uma elevada parte das suas melhorias baseando-se nos dados recolhidos aquando de cada uma das suas interações com o cliente, tentando assim o aprofundamento da relação empresa-cliente, para determinar quais os aspetos de negócio que pode melhorar ou quais os aspetos que o cliente gostaria que fossem resolvidos, por exemplo.

A análise de sentimentos e mineração de opiniões trazem grande benefícios, bem como a utilização dos dados obtidos no estabelecimento de índices de satisfação. Este tipo de índices permite observar os cenários envolventes no que diz respeito à satisfação de consumidores relativamente a um dado produto ou serviço proporcionado por uma empresa ou instituição (McGlohon, Glance and Reiter, 2010). Anteriormente (Secção 1.2) expuseram-se os casos de (Zhan, Loh and Liu, 2009), (Ortigosa, Martín and Carro, 2014), (Anjaria and Guddeti, 2014) e (Breault, Goodall and Fos, 2002), bem como a forma como cada um destes trabalhos demonstrou os benefícios da mineração de opiniões e de sentimentos para disponibilizações intuitivas e eficientes das polaridades existentes e evidenciadas em situações reais. Deste modo, o propósito da aplicação prende-se, sobretudo, com a disponibilização de dados, mais concretamente, índices de satisfação, num cenário real, com temas reais e da sociedade atual, pretendendo-se que o utilizador possa, através da utilização da aplicação, retirar fundamentos consistentes para utilização nos seus processos de tomada de decisão. Além da disponibilização de índices de satisfação relativamente a todo e qualquer tema inserido pelo utilizador, pretende-se também sustentar os resultados obtidos pela análise de dados, demonstrando especificamente e objetivamente o porquê dos resultados apresentados, enumerando os valores avaliativos obtidos para cada análise disponibilizada.

Assim, pode concluir-se que a mineração e análise de opiniões e de sentimentos, implica, em si, um grande redireccionamento de todas as plataformas existentes na *Web*. Hoje em dia, observam-se vários casos de plataformas de marcas de luxo que procuram analisar e constituir perfis de

utilizador, de forma a conseguirem, para assim e respetivamente para cada utilizador específico, apresentar o tipo de produto que o mais possa cativar.

Todas estas extrações e análises de dados definem um novo paradigma em todo e qualquer tipo de mercado existente, uma vez que se observa que a riqueza dos dados gerados a todo e qualquer momento constitui um valor inestimável para a sustentabilidade do progresso de uma empresa ou de um indivíduo.

4.3 Objetivos da Aplicação

Com o desenvolvimento desta aplicação pretende-se disponibilizar um conjunto de dados de análise que proporcione ao utilizador, da forma mais intuitiva possível, algumas perspetivas acerca do modelo de negócio que tem implementado, como uma forma de suporte e auxílio ao seu processo de tomada de decisão. Por outro lado, ao nível do plano individual, fornecer um cenário simples para a prospecção do mercado, no qual o utilizador poderá analisar diferentes produtos, fazendo uma análise exaustiva das várias ofertas existentes de modo a que a sua escolha de compra recaia no produto mais adequado para as suas necessidades.

Desta forma, poderemos dar um contributo bastante válido para melhorar a forma como as decisões são fundamentadas, tanto num âmbito empresarial como forma de decidir o futuro da empresa, tanto em termos empresariais, para decidir o futuro da empresa, como em termos individuais, para auxiliar o utilizador na prospecção de mercado que se propõe a realizar. Além disso, existe um aumento do número de pontos de vista fornecidos sobre o tema pesquisado na aplicação, que são baseados em opiniões reais que refletem a crítica social, seja ela positiva, neutra ou negativa, tanto para uma empresa, que pretende

saber mais sobre o mercado, como para um simples cliente que deseje efetuar um estudo de mercado sobre certo produto antes de adquiri-lo.

Adicionalmente, existe também a possibilidade de visualizar, a nível cronológico, qual o nível de desenvolvimento de certo tema, produto ou serviço. Desta forma, a análise das opiniões sociais sobre certo tema ao longo do tempo, podem ser mais consolidadas e mais bem suportadas relativamente ao tema em questão. De referir que, esta perspetiva cronológica segue de perto os três parâmetros avaliativos do tema: positividade, neutralidade, negatividade. Sendo a quantidade da amostra inserida um valor variável, existe a possibilidade de haver variação da granularidade pretendida para uma dada pesquisa, aumentando ou diminuindo o universo representativo da amostra inserida. Todavia, maioritariamente, espera-se uma maior precisão dos resultados obtidos quanto maior for a amostra extraída.

Para cada um dos parâmetros avaliativos analisados em cada tema, nomeadamente a positividade, a neutralidade e a negatividade, a aplicação permite apresentar uma análise pormenorizada, nomeadamente através de uma *wordcloud*, para cada um dos parâmetros, e, para além disto, o conjunto dos *tweets* analisados que pertencem a cada parâmetro em cada página, respetivamente. Nesse conjunto referem-se ainda os valores de polaridade e subjetividade relativamente a cada *tweet*. Desde modo, pode ocorrer a possibilidade de retirada de conclusões fundamentadas mediante a disponibilização de resultados providenciados na aplicação, sendo que tais resultados são uma referência para uma melhor análise de mercado antes do processo de tomada de decisões, quer seja realizado por parte de uma empresa quer por parte de um agente individual que pretende adquirir determinado produto e/ou serviço.

4.4 A Implementação

A aplicação desenvolvida designa-se por *OpClassify*, uma designação resultante da derivação da junção das palavras *Opinion* e *Classify*, em português, Classificação de Opiniões. O nome foi escolhido tendo por base o raciocínio que foi tido aquando do desenho e da idealização da arquitetura para esta aplicação, que consiste na receção como *input* de um tema em forma de *string* quer seja este do domínio político, social, cultural ou desportivo ou qualquer um outro que crie conteúdo opinativo e passível de ser analisado após extraído da plataforma *Twitter*. A abordagem utilizada no desenvolvimento da aplicação foi a metodologia CRISP-DM, que orientou todo o planeamento da implementação do sistema.

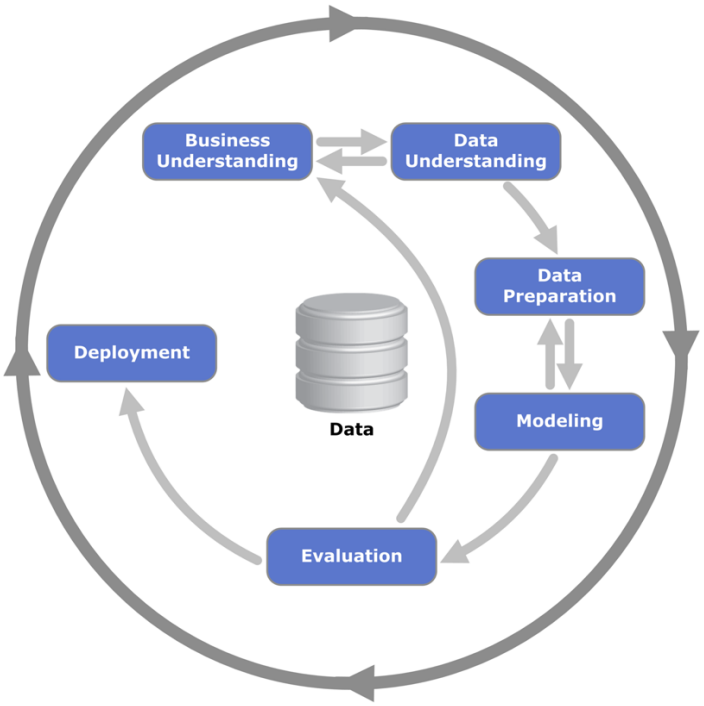


Figura 14: Metodologia CRISP-DM
(Fonte: https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining#/media/File:CRISP-DM_Process_Diagram.png)

Na aplicação, após o utilizador indicar o tema a trabalhar, é executada a análise da polaridade de opiniões que são publicadas relativamente a este mesmo tema inserido inicialmente pelo utilizador. O tema inserido faz despoletar todo o processo de análise de sentimentos e de mineração de opiniões existentes relacionadas com o tema em causa. Como foi dito, os requisitos da aplicação prendem-se com a necessidade do estabelecimento de índices de satisfação através da análise de opiniões e sentimentos. Sendo assim, é necessário a análise de opiniões e sentimentos sobre um dado tema, tema este inserido pelo utilizador da aplicação.

A aplicação foi desenhada tendo em conta toda a arquitetura planeada para o projeto e os seus respetivos objetivos. A linguagem escolhida para a implementação da aplicação foi a *Python* e framework de desenvolvimento adoptada foi a *Django*.

A biblioteca *pandas* também foi utilizada para criar um *dataframe* contendo as frases relativas aos tweets partilhados sobre o assunto inserido. Na coluna *tweets* dispomos as frases consecutivamente, e na coluna *SA* (*Sentiment Analysis*) são dispostos os resultados relativos às respetivas frases, resultados estes obtidos tendo em conta o resultado da avaliação da polaridade executada sobre cada uma das frases.

4.5 A Extração dos Dados

A base da aplicação é sustentada através do processo denominado por *Web scrapping* ou *Web data extraction*. Este processo inicia-se com a extração

de dados a partir de uma fonte de dados, que, neste caso específico, é a fonte de dados associada com a plataforma *Twitter*. O *Twitter* é bastante reconhecido (Russell, 2011)(Chen and Zimbra, 2010)(Patodkar and I.R, 2016) como uma plataforma na qual as opiniões individuais possuem um papel central naquilo que pode servir de base para retirar ilações sobre perspetivas sociais que abordam os mais variados assuntos ocorridos em todo o mundo e nas mais variadas vertentes de problemáticas socioculturais. Nesta plataforma (Dokoohaki *et al.*, 2015)(Nielsen, 2011) tomam parte muitas figuras centrais no plano político mundial e, de forma mais particular, do plano político de cada país. O mesmo acontece no plano desportivo e no plano social.

A inserção de um tema sociocultural, desportivo ou político na página principal da aplicação despoleta a extração de *tweets* existentes sobre esse mesmo tema.

Este é, pois, o ponto de partida para a realização de um processo de análise, tendo em conta o respetivo tema inserido.

```
def index(request):
    print ("REQUEST")
    tema=request.POST.get('theme', None)

    if (tema):
        print (tema)
        auth=OAuthHandler(consumer_key, consumer_secret)
        auth.set_access_token(access_token, access_secret)

        api=tweepy.API(auth)

        tweets=list()

        #for status in tweepy.Cursor(api.home_timeline).items(10):
        #for status in tweepy.Cursor(api.search, q='#trump', rpp=100, count=200, tweet_mode="extended").items(20):
        for status in tweepy.Cursor(api.search, q=tema, rpp=100, count=200, tweet_mode="extended").items(20):
            # Process a single status
            #print(status)

            #print(json.dumps(status, indent=4)) # pretty-print

            #process_or_store(status._json)
            tweets.append(status.full_text)
            print ("###################\n")
            print(preprocess(status.full_text))
        #Analyze the text#

        print ("Lista de Tweets sobre o assunto inserido:")
        print (tweets)
        # We create a pandas dataframe as follows:
        dataTweets = pd.DataFrame(data=[tweet for tweet in tweets], columns=['Tweets'])
        # We display the first 10 elements of the dataframe:
        display(dataTweets.head(10))

        dataTweets['SA'] = np.array([ analize_sentiment(tweet) for tweet in dataTweets['Tweets'] ])

        # We display the updated dataframe with the new column:
        display(dataTweets.head(10))

        # We construct lists with classified tweets:

        pos_tweets = [ tweet for index, tweet in enumerate(dataTweets['Tweets']) if dataTweets['SA'][index] > 0]
        neu_tweets = [ tweet for index, tweet in enumerate(dataTweets['Tweets']) if dataTweets['SA'][index] == 0]
        neg_tweets = [ tweet for index, tweet in enumerate(dataTweets['Tweets']) if dataTweets['SA'][index] < 0]

        # We print percentages:

        print("Percentage of positive tweets: {}%".format(len(pos_tweets)*100/len(dataTweets['Tweets'])))
        print("Percentage of neutral tweets: {}%".format(len(neu_tweets)*100/len(dataTweets['Tweets'])))
        print("Percentage de negative tweets: {}%".format(len(neg_tweets)*100/len(dataTweets['Tweets'])))

        #Analyze the text#
```

Figura 15: Visualização do ficheiro views.py

Na Figura 15 podemos ver a função *index* que recebe o primeiro pedido feito pela aplicação (em resposta ao *url index* – "*/*"). Neste processo, em primeiro lugar, é obtida a *authentication* necessária junto da plataforma Twitter com um objeto *OAuthHandler*, para que seja possível, desta forma, inicializar uma instância da conexão à *API* do *Tweepy* (Novalić, 2013)(Raut, 2016)(Mishra, Khandelwal and Mishra, 2016) [12]. Através do *Tweepy* é depois processada a extração de 200 tweets sobre o tema recebido na aplicação como *input*. Após tal ter acontecido, o atributo *full_text* de cada *tweet* é armazenado para posterior análise.

4.6 O Pré-processamento dos Dados

A fase de pré-processamento dos dados é uma etapa crucial para que o sucesso seja atingido pelo sistema de análise de dados em relação aos resultados por si produzidos. A existência de dados incompletos, dados mal formatados e a ocorrência de não-uniformização de campos (e.g. numa base de dados o País de uma pessoa com nacionalidade portuguesa pode ser escrito como Portugal, PT, Port, Prtgal), são exemplos que demonstram como o sucesso de um sistema de análise de dados pode, à partida, ficar comprometido. O pré-processamento dos dados tem em atenção este tipo de problemas e serve como resposta à necessidade da resolução dos mesmos.

(Dey and Haque, 2009) afirmaram que a fase de pré-processamento dos dados textuais é um estágio basilar de todo o processo de análise. Esta permite-

[12] *Tweepy* é uma biblioteca em *Python* que providencia o acesso à *API* do *Twitter*.

nos diferenciar os vários tipos de elementos contidos na frase que se encontra a ser analisada. Mais especificamente, no caso da análise de dados textuais abordado neste trabalho, o pré-processamento dos *tweets* extraídos é essencial. Um simples *tweet* pode conter imensos elementos que dificultem o processo de análise textual, como por exemplo, caracteres especiais ou símbolos, que até por engano, façam parte do *tweet*. De modo a responder a esta necessidade de pré-processamento dos tweets extraídos para análise na aplicação implementada, foi codificada a função *preprocess*. A função *preprocess* pode ver-se na Figura 16. Esta função é responsável por remover caracteres especiais, *links* e outro tipo de palavras que dificultam a análise das frases. Para atingir este fim, a função preprocess faz uso da definição de expressões regulares. Esta mesma função é invocada sobre todas as frases que vão ser analisadas, de modo a executar o pré-processamento dos dados textuais a analisar.

```
import re

emoticons_str = r"""
    (?:
        [:=;] # Eyes
        [oO\-]? # Nose (optional)
        [D\)\]\(\]/\\OpP] # Mouth
    )"""

regex_str = [
    emoticons_str,
    r'<[^>]+>', # HTML tags
    r'(?:@[\w_]+)', # @-mentions
    r"(?:\#+[\w_]+[\w\'_\-]*[\w_]+)", # hash-tags
    r'http[s]?://(?:[a-z]|[0-9]|[$-_@.&amp;+]|[!*\(\),]|(?:%[0-9a-f][0-9a-f]))+', # URLs

    r'(?:(?:\d+,?)+(?:\.?\d+)?)', # numbers
    r"(?:[a-z][a-z'\-_]+[a-z])", # words with - and '
    r'(?:[\w_]+)', # other words
    r'(?:\S)' # anything else
]

tokens_re = re.compile(r'('+'|'.join(regex_str)+')', re.VERBOSE | re.IGNORECASE)
emoticon_re = re.compile(r'^'+emoticons_str+'$', re.VERBOSE | re.IGNORECASE)

def tokenize(s):
    return tokens_re.findall(s)

def preprocess(s, lowercase=False):
    tokens = tokenize(s)
    if lowercase:
        tokens = [token if emoticon_re.search(token) else token.lower() for token in tokens]
    return tokens
```

Figura 16: Visualização do ficheiro views.py – funções preprocess, tokenize, e elemento regex.

A função *preprocess* recebe como *input* o atributo *full_text* de cada *tweet* e executa o *parsing* de cada palavra, tendo em conta durante esse mesmo

parsing as etiquetas *HTML* e as menções do tipo *@-mentions*, *hashtags*, *URLs*, números, e palavras com hífens e apóstrofes.

```
def clean_tweet(tweet):
    '''
    Utility function to clean the text in a tweet by removing
    links and special characters using regex.
    '''
    return ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)", " ", tweet).split())
```

Figura 17: Visualização do ficheiro views.py - função clean_tweet.

Na Figura 17 pode observar-se a função *clean_tweet*. Esta tem como função "limpar" o *tweet* prestes a ser invocado como argumento da função *analize_sentiment*. Todos os links e caracteres especiais são filtrados através do uso de expressões regulares.

Através do uso destas funções, o pré-processamento dos tweets extraídos é assegurado, permitindo a obtenção de uma boa qualidade de dados para análise na etapa seguinte do *workflow* da aplicação.

4.7 A Análise dos Dados Recolhidos

Após o pré-processamento dos dados recolhidos, o processo continua com a execução de uma função sobre os textos dos *tweets* pré-processados. Nesta parte é utilizada a biblioteca *textblob* para análise da polaridade dos textos, que poderá assumir um de três resultados possíveis: positivo, neutro e negativo.



```
51              #process_or_store(status._json)
52              tweets.append(status.full_text)
53              print ("##################\n")
54              print(preprocess(status.full_text))
55          #Analyze the text#
56
57          print ("Lista de Tweets sobre o assunto inserido:")
58          print (tweets)
59          # We create a pandas dataframe as follows:
60          dataTweets = pd.DataFrame(data=[tweet for tweet in tweets], columns=['Tweets'])
61          # We display the first 10 elements of the dataframe:
62          display(dataTweets.head(10))
63
64          dataTweets['SA'] = np.array([ analize_sentiment(tweet) for tweet in dataTweets['Tweets'] ])
65
66          # We display the updated dataframe with the new column:
67          display(dataTweets.head(10))
68
69          # We construct lists with classified tweets:
70
71          pos_tweets = [ tweet for index, tweet in enumerate(dataTweets['Tweets']) if dataTweets['SA'][index] > 0]
72          neu_tweets = [ tweet for index, tweet in enumerate(dataTweets['Tweets']) if dataTweets['SA'][index] == 0]
73          neg_tweets = [ tweet for index, tweet in enumerate(dataTweets['Tweets']) if dataTweets['SA'][index] < 0]
74
75          # We print percentages:
76
77          print("Percentage of positive tweets: {}%".format(len(pos_tweets)*100/len(dataTweets['Tweets'])))
78          print("Percentage of neutral tweets: {}%".format(len(neu_tweets)*100/len(dataTweets['Tweets'])))
79          print("Percentage de negative tweets: {}%".format(len(neg_tweets)*100/len(dataTweets['Tweets'])))
80
81          #Analyze the text#
82
```

Figura 18: Visualização do ficheiro views.py - Display dos resultados dos dados analisados.

Seguidamente, e recorrendo ao uso da biblioteca pandas é criado um *dataframe* com a informação extraída, integrando duas colunas: a "*Tweets*", contendo o texto em si constituinte do respetivo tweet extraído e correspondente aos elementos analisados, e a "*SA*" - *Sentiment Analysis*, que conterá, após calculado, o resultado obtido relativamente à polaridade do tweet respetivo. Na linha 60 cria-se a coluna *"Tweets"* no dataframe, inserindo cada *tweet* extraído na mesma. Na linha 64 adiciona-se a coluna *"SA"* – *Sentiment Analysis* ao *dataframe* criado acima na linha 60. Para cada texto pré-processado, a função *analize_sentiment* é invocada, tendo como argumentos esses mesmos elementos do tweet inicialmente extraído.

Para melhor ilustrar o proveito que pode ser retirado da biblioteca *TextBlob*, dispõe-se, na Figura 19, um exemplo de algumas das capacidades benéficas da biblioteca relativamente à sua utilização para fins de análise de dados textuais.

```
>>> b=TextBlob("The movie was fantastic.")
>>> b.tags
[('The', u'DT'), ('movie', u'NN'), ('was', u'VBD'), ('fantastic', u'JJ')]
>>> b.noun_phrases
WordList([])
>>> b.words
WordList(['The', 'movie', 'was', 'fantastic'])
>>> b.sentiment
Sentiment(polarity=0.4, subjectivity=0.9)
```

Figura 19: Exemplo da análise de polaridade usando a biblioteca TextBlob.

Como demonstrado na Figura 19, a biblioteca *TextBlob* quando aliada à linguagem Python, formam um conjunto de ferramentas bastante eficaz na execução de mineração de opiniões sobre qualquer tipo de texto.

Na linha 1 da Figura 19 pode-se observar a criação de um objeto *TextBlob* com a frase *"The movie was fantastic."*. No passo seguinte, na linha 3, observa-se a etiquetagem da frase em relação às etiquetas *POS* (*Part-Of-Speech*). Através do código *b.words*, na linha 6, é retornada uma lista das palavras que constituem as unidades elementares da frase *"The movie was fantastic."*. Por fim, com o código *b.sentiment* obtém-se o grau de polaridade e o grau de subjetividade para a frase em questão. Este exemplo, ainda que pequeno, serve para demonstrar a utilidade da biblioteca TextBlob na análise dos tweets extraídos, de modo a que, se pudessem obter resultados consolidados para a disponibilização de resultados que se apresentam na secção seguinte.

4.8 Disponibilização dos Resultados

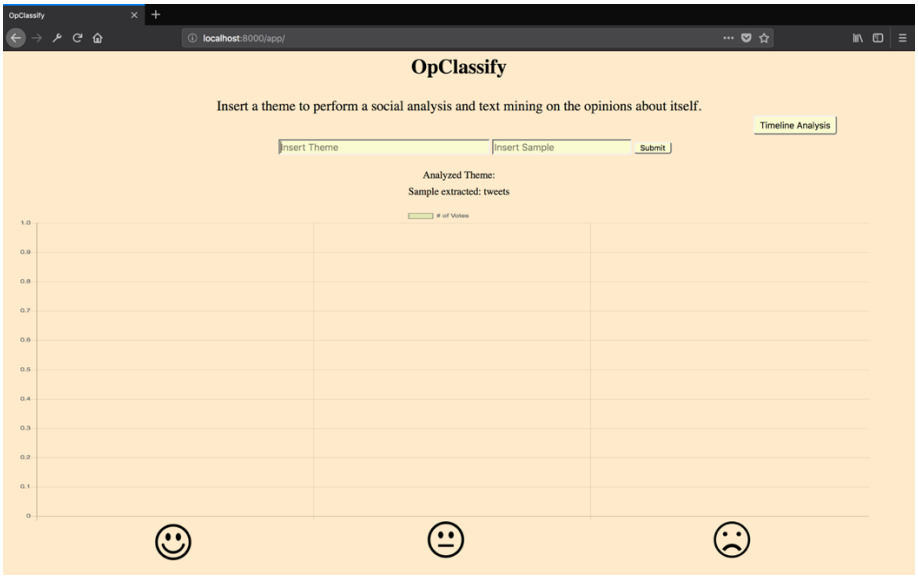


Figura 20: Página inicial da aplicação OpClassify

Na Figura 20 é possível observar a página inicial da aplicação desenvolvida, com um input que foi implementado para fazer a inserção do tema que despoleta todo o processo de análise de sentimentos e de mineração de opiniões partilhadas na plataforma Twitter sobre o tema em questão. Além deste *input*, existe também um outro *input* para a inserção de um inteiro que representa a amostra a ser recolhida sobre esse tema em que estamos a trabalhar, ou seja, o número de *tweets* que o utilizador pretende que sejam recolhidos sobre esse tema, o que permite fazer a afinação dos possíveis resultados a obter junto da aplicação, aumentando ou diminuindo o nível de granularidade da pesquisa do tema.

Se omitirmos o inteiro representativo da quantidade da amostra a ser posteriormente recolhida, apenas serão recolhidos dez *tweets* sobre o tema inserido. A título de exemplo insere-se o tema "*brexit*", com uma amostra de 100 na pesquisa realizada na aplicação (Figura 21).

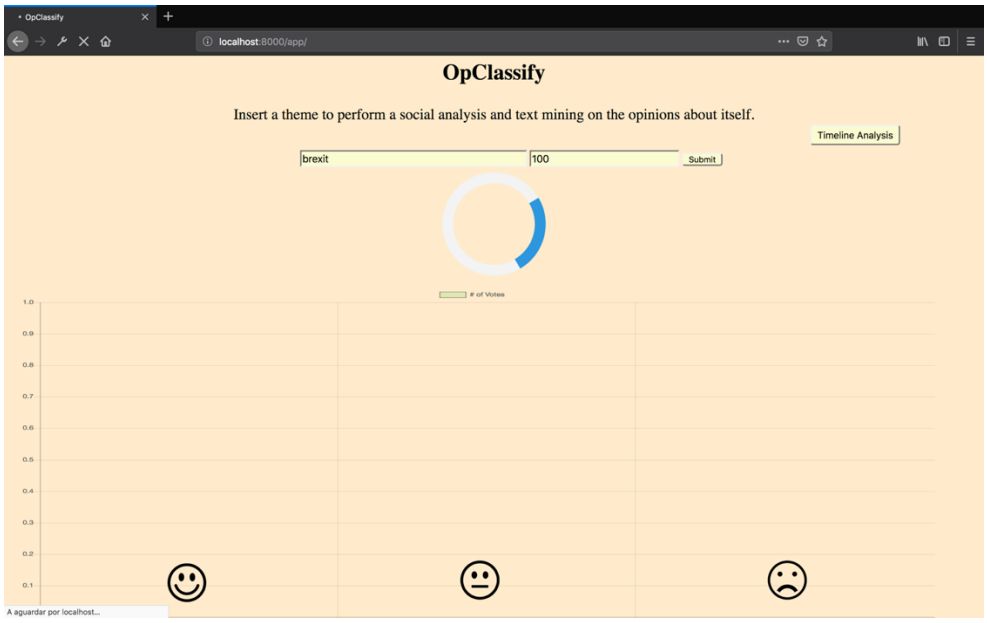


Figura 21: Inserção do tema "brexit" com amostra 10 e loader do sistema em execução.

Após a inserção do tema e da respetiva amostra podemos verificar o aparecimento de um *loader* indicativo da execução momentânea do processo de extração e análise dos dados extraídos (Figura 21).



Figura 22: Exemplo de pesquisa do tema "brexit" no OpClassify.

No gráfico existente na Figura 22 revelam-se os níveis de positividade, de neutralidade, e de negatividade, que estão etiquetados com ☺, 😐 e ☹, respetivamente, em termos percentuais e com o detalhe da cor associada a cada construção gráfica de percentagem relacionada com a polaridade da opinião, sendo a positividade denotada com a cor verde, a neutralidade com a cor amarela e a negatividade com a cor vermelha. Ao selecionarmos o símbolo representativo da positividade (☺), a aplicação disponibiliza uma nova página com uma análise dos *tweets* que foram considerados positivos, conjuntamente com o seu resultado relativamente aos fatores polaridade e subjetividade. No topo dessa página é possível observar uma *wordcloud* representativa do conjunto de palavras mais utilizadas nos *tweets* que foram considerados positivos – todos estes factos podem ser vistos na Figura 23.

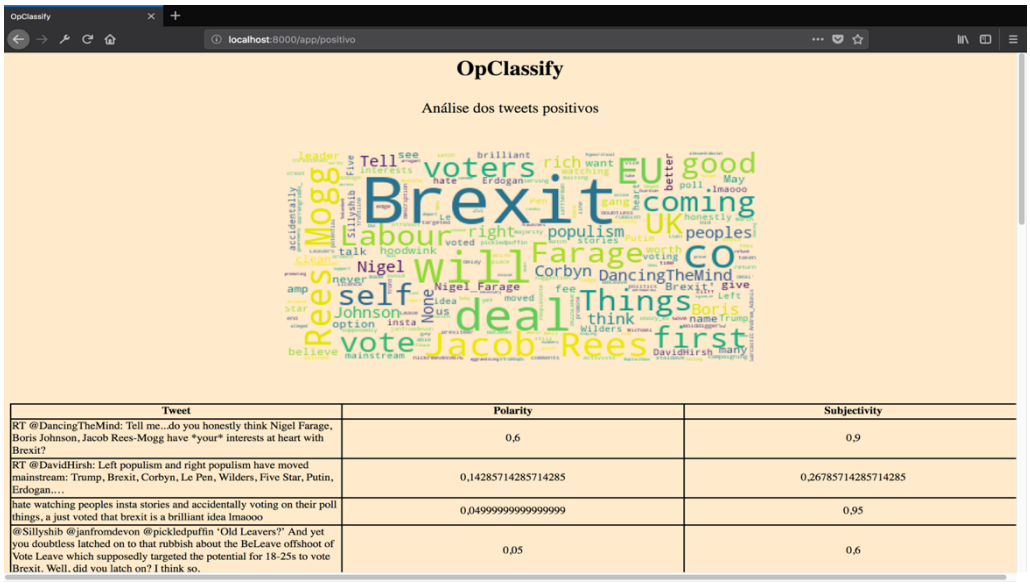


Figura 23: Análise dos Tweets Positivos relativamente à pesquisa sobre o tema "brexit".

Se o utilizador selecionar a opção relativa à neutralidade (😐), a aplicação disponibiliza uma nova página que representa a análise dos *tweets* que foram considerados neutros, conjuntamente com o seu resultado relativamente aos fatores de polaridade e de subjetividade. No topo dessa página é possível observar uma *wordcloud* representativa do conjunto de palavras mais utilizadas nos *tweets*, cuja polaridade foi considerada neutra (Figura 24).

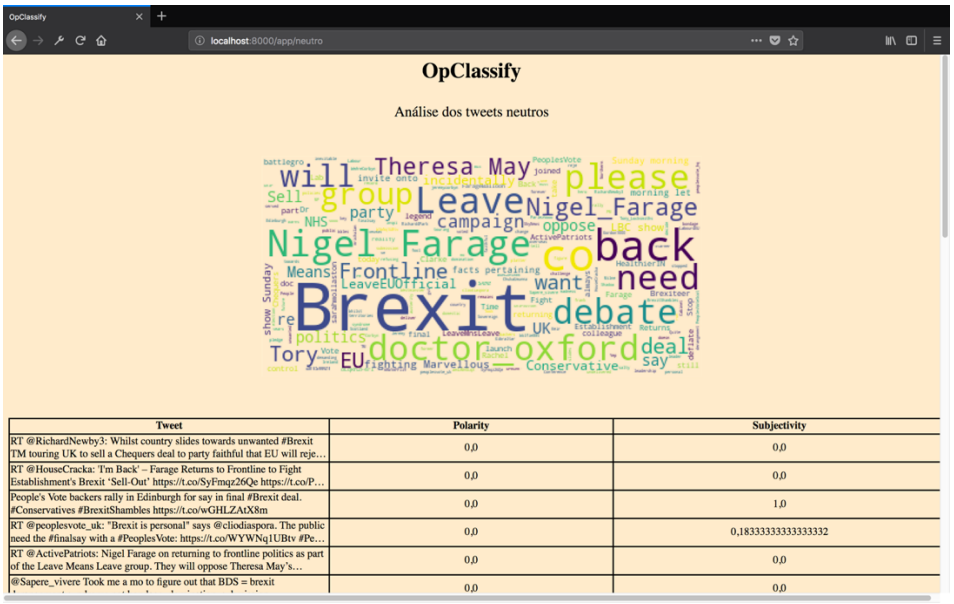


Figura 24: Análise dos Tweets Neutros relativamente à pesquisa sobre o tema "brexit".

Por sua vez, se o utilizador selecionar a opção relativa à negatividade (☹), a aplicação disponibiliza, uma vez mais, uma nova página que revela a análise dos *tweets* que foram considerados negativos, conjuntamente com o seu resultado relativamente aos fatores de polaridade e de subjetividade. No topo dessa página é possível observar, tal como no caso anterior, uma *wordcloud* representativa do conjunto de palavras mais utilizadas nos *tweets*, cuja polaridade foi considerada negativa (Figura 25).

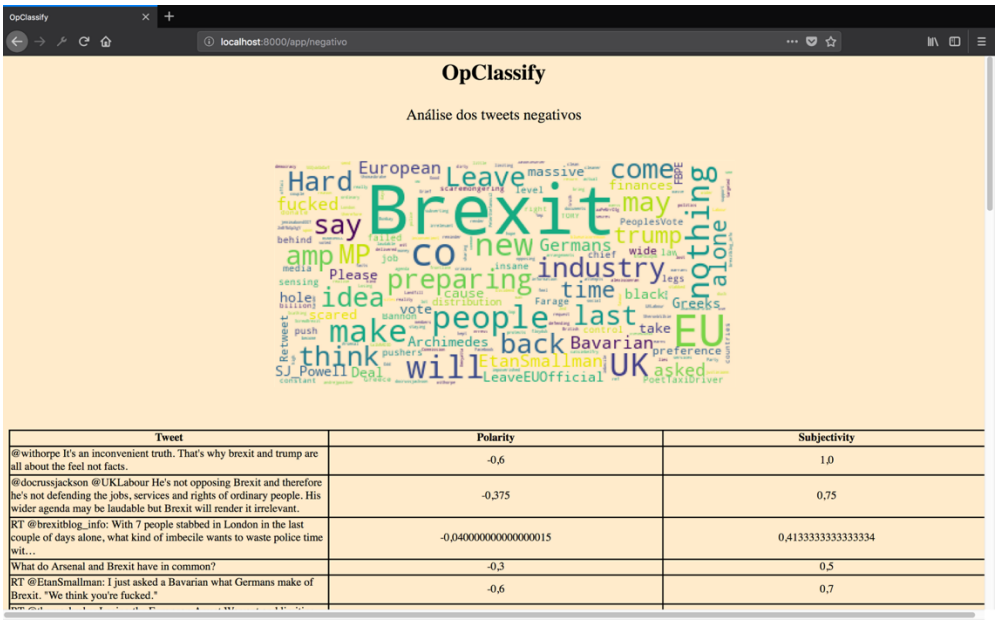


Figura 25: Análise dos Tweets Negativos relativamente à pesquisa sobre o tema "brexit".

Voltando à página principal da aplicação, se o utilizador selecionar a opção Timeline Analysis, este é redirecionado para uma página que permite uma análise cronológica da evolução da polaridade sobre um mesmo tempo ao longo do tempo (Figura 26). Um exemplo da aplicação destes resultados pode ser observado na Figura 26, na qual está representado um dado intervalo de anos e, para cada um dos meses considerados, se apresenta um valor denominado de classificação *opclassify* ou *opclassification*. Este valor é, fundamentalmente, um índice de popularidade do tema inserido, ou seja, um indicador calculador com base nos níveis de positividade, neutralidade e negatividade do tema inserido.

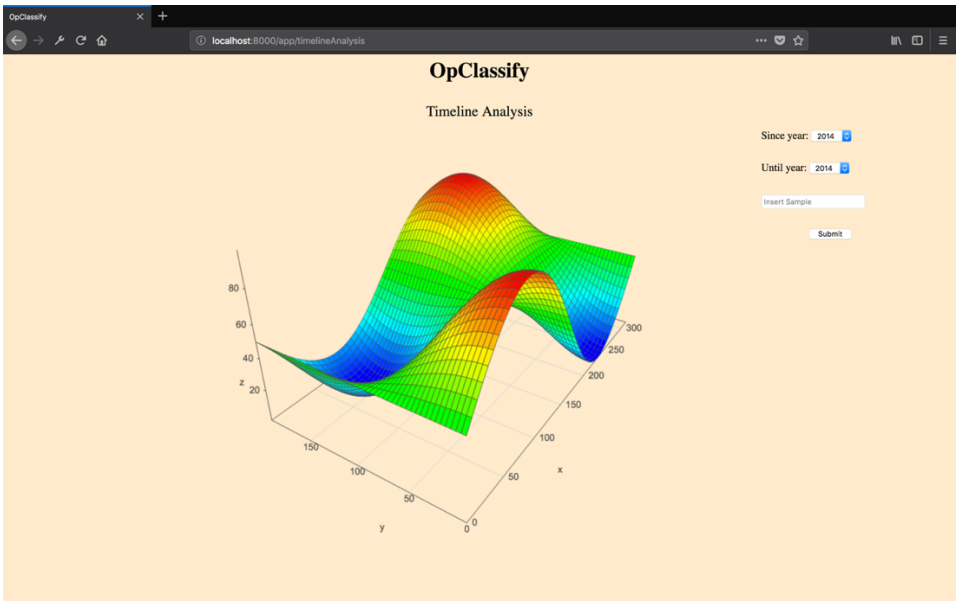


Figura 26: Arquitetura implementada para linha cronológica baseada na análise por mês/ano.

Este índice corresponde ao valor percentual atribuído ao tema, numa escala de 0 a 100, segundo as opiniões sociais acerca dele. Basicamente, é uma resposta à seguinte questão: "Numa escala de 0 a 100 e segundo as opiniões sociais sobre o tema X, qual o valor atribuído à opinião social sobre o tema X?". Dos possíveis valores que o índice *opclassification* pode tomar, 0 significa que em torno do tema apenas existem comentários negativos, 50 significa que o tema resultado em diversas opiniões divergentes, e 100 representa um valor exclusivo de positividade presentes nos comentários alusivos ao tema em causa.

A *opclassification* para um dado tema, denotada por OP(tema), pode ser calculada da seguinte forma:

$$OP(tema) = índice\ de\ positividade + índice\ de\ neutralidade \times 0.5$$

Suponhamos a título de exemplo que utilizámos o tema *"brexit"* para análise, com uma amostra de 100 *tweets*. Com a inserção desse tema na aplicação, obtivemos resultados com 40% para o fator positividade, 20% para o

fator neutralidade, e 40% para o fator negatividade. Com estes dados, podemos afirmar que:

$$OP(\text{brexit}) = 40 + 20 \times 0.5 = 40 + 10 = 50\%$$

O que significa que a *opclassification* para o tema *"brexit"* é de 50%.

Durante a implementação da aplicação foi verificado que para se aceder a *tweets* realizados num dado ano em específico é necessário fazer a aquisição de um acesso *premium* à *API* do *Twitter*. Com base nisso, decidiu-se utilizar o acesso aos dados disponibilizados através de um plano gratuito, o que apenas permite o acesso aos *posts* realizados no *Twitter* nos últimos 7 dias. Assim, qualquer input que queiramos definir restringe-se, obviamente, à quantidade de dias passados que se pretende incluir na análise a realizar. Ou seja, tendo como exemplo o tema "brexit", com uma amostra de 100 elementos para análise e definindo como 3 o número de dias passados a analisar, foi possível realizar a análise que a Figura 27 revela.

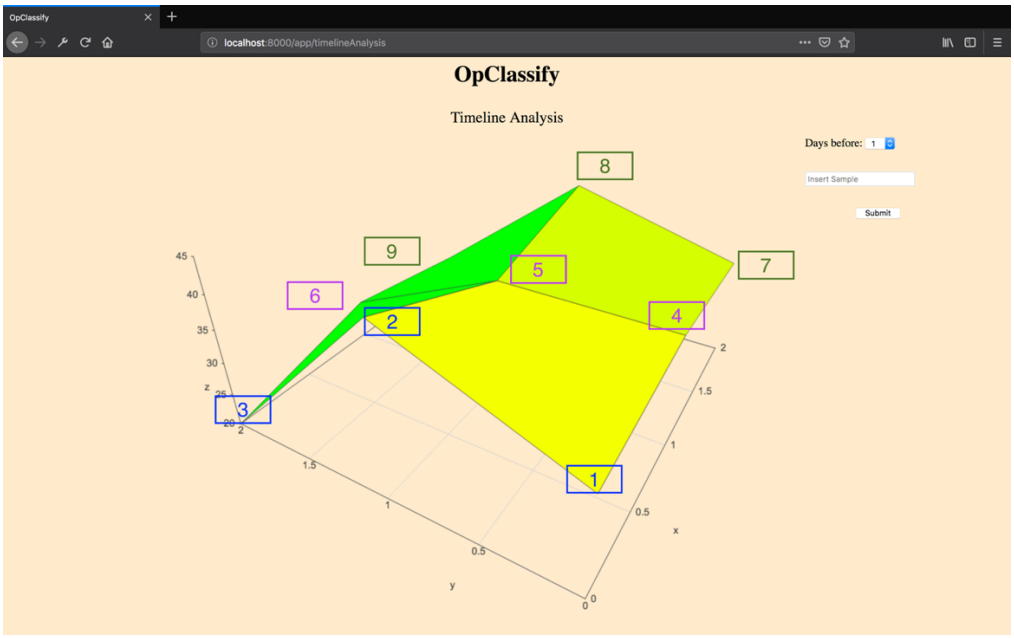


Figura 27: 1 - Índice de negatividade do dia anterior; 2 - Índice de neutralidade do dia anterior; 3 - Índice de positividade do dia anterior; 4 - Índice de negatividade de 2 dias anteriores; 5 - Índice de neutralidade de 2 dias anteriores; 6 - Índice de positividade de 2 dias anteriores; 7 - Índice de negatividade de 3 dias anteriores; 8 - Índice de neutralidade de 3 dias anteriores; 9 - Índice de positividade de 3 dias anteriores.

Na Figura 27 podem-se observar 3 retas perpendiculares ao eixo das abcissas. Cada uma dessas retas representa um dia de dados, enquanto que as 3 retas perpendiculares ao eixo das coordenadas representam a evolução do fator negatividade nos pontos (1, 4, 7), do fator neutralidade (2, 5, 8), e do fator negatividade (3, 6, 9). Estes tuplos, com três elementos, referem-se, respetivamente, ao dia imediatamente antes do atual e ao número de dias passados até ao limite inserido.

4.9 Comparação de Resultados

A comparação de resultados bem estruturada exige um igualdade de condições na avaliação dos resultados obtidos por cada uma das parte envolvidas no processo. Nesta comparação de resultados, e, para que fosse estabelecida a maior igualdade possível entre as diferentes perspetivas em análise, o código foi alterado de modo a que apenas se procedesse à extração dos dados uma vez apenas. Após isso, e sobre os dados recolhidos, executou-se a análise dos mesmos através de cada uma das abordagens em avaliação.

Como forma de contextualizar a existência das duas perspetivas em análise, adianta dizer que a biblioteca que utilizámos na análise das opiniões e sentimentos extraídos do *Twitter* foi a *textblob*[13], tal como é referido nas secções anteriores relativas à implementação da aplicação. Esta biblioteca, escrita em *Python*, permite fazer o processamento de dados textuais. Através de um dos seus módulos, nomeadamente o *texblob.sentiments* acedemos a duas implementações de análise de sentimentos, em particular, através do analisador padrão baseado no módulo de mineração *Web* denominado padrão e o modelo de classificadores *Naive Bayes*[14].

[13] https://textblob.readthedocs.io/en/dev/#

[14] Modelo classificador baseado no teorema de Naive Bayes (probabilidade de A acontecer sabendo que B aconteceu).



Figura 28: Módulo de mineração Web padrão (Fonte: https://www.clips.uantwerpen.be/pattern).

O modelo Naive Bayes é um modelo suportado pelo *toolkit* de linguagem natural NLTK. Este toolkit providencia acesso a vários recursos léxicos e a mais de 50 corpus linguísticos, além de que possui também um conjunto de bibliotecas especialmente orientado para o processamento de texto.[15] Além disso, este modelo foi treinado baseado num corpus de críticas cinematográficas. Assim, a análise dos *tweets* extraídos através da aplicação foi realizada através da utilização destes dois métodos.

O processo de comparação realizado teve como base a análise de cinco temas diferentes, cada um deles com uma amostra de 10000 *tweets*. Os temas escolhidos para análise foram os seguintes:

- Brexit

  Os britânicos decidiram em referendo a saída da Inglaterra da União Europeia. Esta decisão poderá colocar em causa o futuro da União Europeia.

- Facebook;

  Na sociedade atual, o Facebook é uma aplicação que está a mudar o modo de comunicar a nível global. A sua utilização conquistou todas as

[15] De acordo com o referido em http://www.nltk.org.

faixas etárias da sociedade, e, por isso, um elevadíssimo número de utilizadores. Importa, por isso, perceber se, depois de todas as notícias negativas sobre o modo como a empresa lidou com a venda de dados pessoais, o comportamento dos utilizadores se alterou e, desta forma, a opinião que têm sobre o *Facebook*.

- Corruption;

Sempre a corrupção foi uma realidade que acompanhou a história do Homem. Simultaneamente, foi e é, tema constante nas notícias do dia-a-dia. Sendo propósito anunciado dos poderes políticos combater a corrupção, pretende-se perceber a opinião dos cidadãos sobre tema tão complexo e importante para a vida da sociedade.

- Ryanair;

Esta empresa de aviação veio alterar o paradigma do sector. As viagens de baixo custo introduziram uma nova forma de estar no sector das viagens aéreas. Conquistou muitos aderentes mas nem todos se rendem a um serviço que ainda merece muitas críticas. Sobretudo, ao nível da qualidade do serviço que presta e da relação laboral com os seus colaboradores.

- McDonald's.

Desde a sua criação em 1940 nos Estados Unidos, a cadeia de fast food não pára de conquistar o Mundo. A forma inovadora como se posiciona no mercado, contrasta com polémicas diversas que põem em causa a forma como a McDonald's faz chegar as suas propostas gastronómicas aos consumidores.

A cada um dos temas referidos aplicámos cada um dos modelos de análise anteriormente referidos.

O processo de comparação iniciou-se com a inserção do primeiro tema ("Brexit") selecionando uma amostra de 10000 *tweets*. Aplicámos o mesmo processo de selecção a todos os outros temas, com os mesmos parâmetros. Como é sabido, alguns destes temas suscitam divisão em termos de opinião social. Como tal a análise de cada um deles, é, naturalmente, complexa.

Outro fator que pode afetar os resultados são os tweets especificamente recolhidos para amostra. Por exemplo, para determinado tema cuja polaridade é, socialmente, tida como maioritariamente negativa, os resultados podem mostrar-se, em maior parte, positivos ou neutros. Tal facto pode dever-se à amostra ser reduzida e os *tweets* extraídos, ocasionalmente, serem positivos ou neutros. O mesmo pode suceder em situações contrárias.

Além deste facto, note-se ainda a existência de uma margem de erro variável para a precisão do modelo, que devido a expressões que contenham ironia, ou a uma má análise, podem não ser alta.

A extração dos dados foi realizada a partir dos *tweets* mais recentes para os mais antigos. Os resultados que se obtiveram para cada tema analisado são apresentados nas secções seguintes. Os resultados obtiveram-se a partir da aplicação de cada um dos modelos de análise selecionados, nomeadamente: o modelo padrão e o modelo *Naive Bayes*.

4.9.1 O Tema "Brexit"

A saída do Reino Unido da União Europeia foi e é um tema ainda bastante discutido socialmente. No próprio país em causa é um tema que gera grande

controvérsia com muitas partes a favor, mas também com bastantes partes contra. Assim, dada a relevância que possui socialmente, o "Brexit" foi um dos temas escolhido para análise. De facto, os resultados obtidos espelham, nos dois métodos, a controvérsia existente socialmente relativamente às opiniões relacionadas com o tema. Vejamos então os resultados.

- Modelo padrão

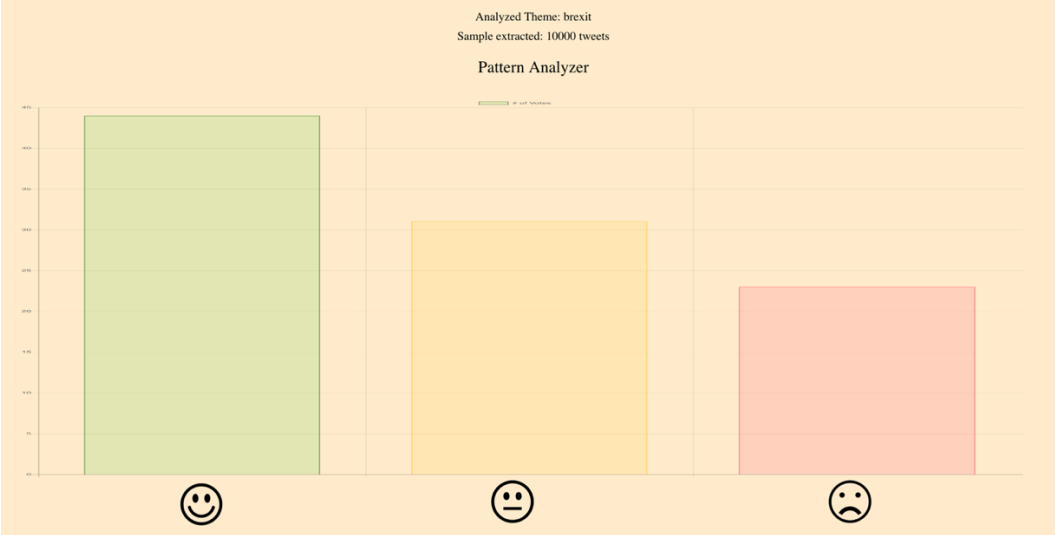


Figura 29: Análise padrão do tema "brexit".

Através da análise pelo modelo padrão (Figura 29) verificou-se que, em relação ao tema *"brexit"*, existe 44.9% de positividade, 31.56% de neutralidade e 23.54% de negatividade. Os valores apresentados revelam uma grande controvérsia de opiniões, com um grande valor percentual atribuído à negatividade. Isto comprova que, realmente, este é um tema fraturante da sociedade britânica, que após o referendo, ainda debate as consequências da saída da União Europeia.

- Modelo Naive Bayes

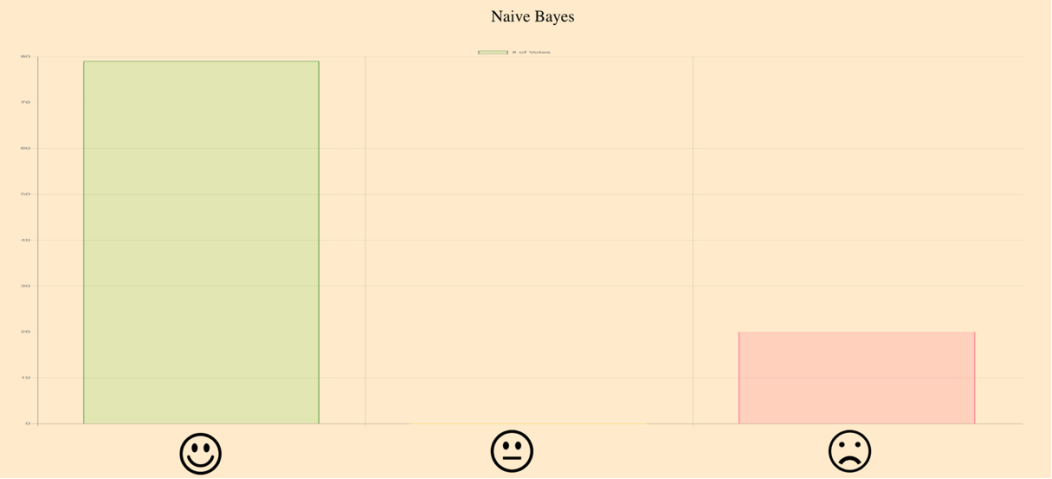


Figura 30: Análise Naive Bayes do tema "brexit".

Através da análise pelo modelo Naive Bayes (Figura 30) identificou-se que, em relação ao tema *"brexit"*, existe 79.16% de positividade, 0.34% de neutralidade, e 20.5% de negatividade. Nesta análise verifica-se que, em contraste com a análise realizada pelo modelo padrão, a percentagem de tweets negativos foi mais baixo enquanto que o valor percentual relativo à positividade foi maior.

4.9.2 O Tema "Facebook"

Como empresa, o Facebook tem tido bastante trabalho para promover a contínua confiança dos utilizadores na sua rede social. As notícias que circularam relativamente ao acesso aos dados gerados pela plataforma por parte de empresas para fins comerciais abalou a imagem pública da empresa perante os seus utilizadores. Devido a isso, o CEO da

Facebook foi praticamente obrigado a justificar-se perante vários governos.

Neste caso que selecionámos para análise, verificámos que ambos os modelos refletem significativamente o estado atual da imagem pública da empresa Facebook. Tendo em conta a amostra de 10000 tweets que utilizámos, os valores apresentados são razoavelmente demonstrativos da reticência existente em relação à empresa. Tal é comprovado pelos valores de 37.31% e 13.8%, respetivamente referentes à neutralidade e negatividade no modelo padrão, e de 2.22% e 31.07%, respetivamente referentes à neutralidade e negatividade no modelo Naive Bayes.

- Modelo padrão

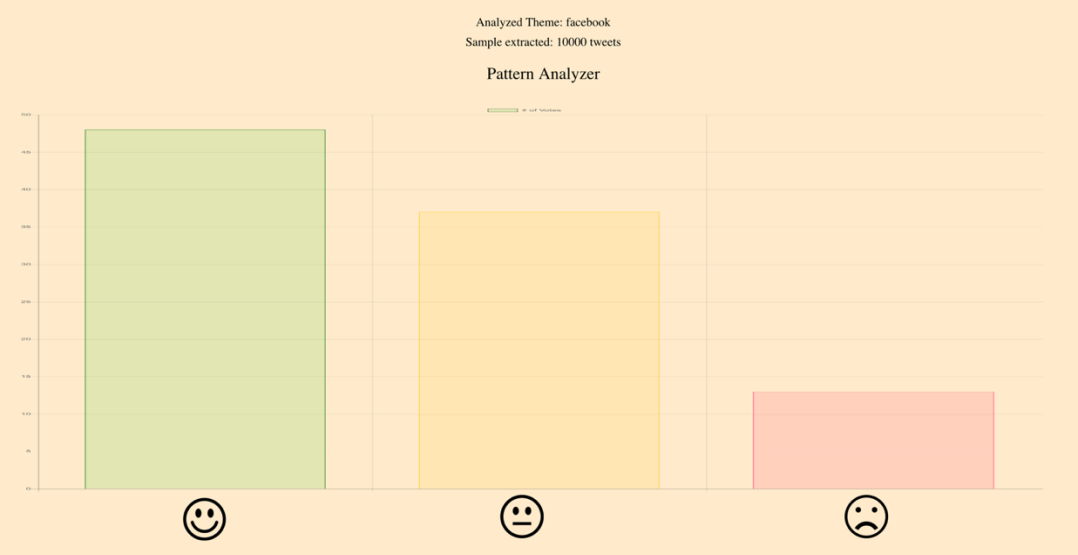


Figura 31: Análise padrão do tema "facebook".

Através da análise pelo modelo padrão observou-se que, em relação ao tema *"facebook"*, existe 48.89% de positividade, 37.31% de neutralidade e 13.8% de negatividade. Estes valores retratam, razoavelmente, o cenário atual relativamente à opinião social existente sobre o Facebook. Muitas pessoas, após o choque das notícias de venda

de dados pessoais não autorizada, por parte do Facebook a outras empresas, questionaram as verdadeiras intenções da empresa. Além disso, colocam em causa a credibilidade da empresa, e que obrigaram à assunção de compromissos públicos sobre o respeito pelos dados pessoais dos utilizadores. Em resultado de tudo isto, a empresa foi fortemente prejudicada nas acções cotadas em bolsa.

- Modelo Naive Bayes

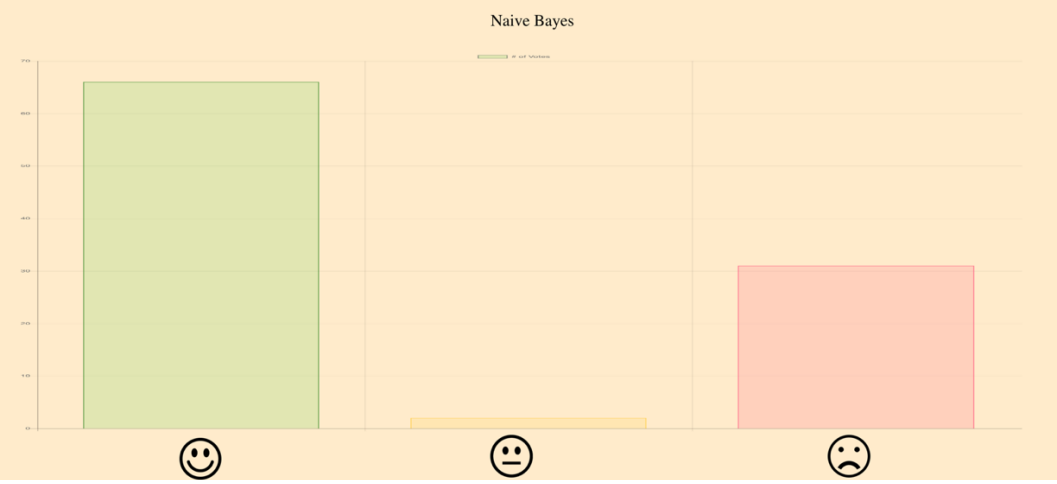


Figura 32: Análise Naive Bayes do tema "facebook".

Através da análise pelo modelo Naive Bayes notou-se que, em relação ao tema *"facebook"*, existe 66.71% de positividade, 2.22% de neutralidade, e 31.07% de negatividade. Os resultados obtidos por este modelo revelam que os utilizadores do Facebook têm posições muito claras sobre o tema. Ainda assim, fica também patente que as opiniões favoráveis são ligeiramente acima do dobro das negativas.

4.9.3 O Tema “Corruption”

Nos dias de hoje, os casos de corrupção são vulgarmente divulgados através dos vários meios de comunicação. Como sabemos, a frequência de ocorrência destes casos é, infelizmente, bastante alta. Como tal, decidimos incluir este tema no nosso processo de análise. À partida, este é um tema claramente negativo, no qual se espera uma extração de um conjunto significativo de tweets, que podem ser considerados como tendo uma análise que resulta em polaridade negativa. Na análise executada com o modelo padrão é possível observar que, de facto, a maioria dos tweets extraídos se situa, segundo este modelo, entre o neutro e o negativo. É, de facto, um resultado que ilustra a realidade da opinião social referente a este tema.

- Modelo padrão



Figura 33: Análise padrão do tema "corruption".

Através da análise realizada pelo modelo padrão, os resultados obtidos revelaram que, em relação ao tema *"corruption"*, os valores

obtidos são de 22.18% de positividade, 47.53% de neutralidade e 30.29% de negatividade. Estes valores indicam que, muito perto de metade dos *tweets* recolhidos não tomam posição sobre o tema. Eventualmente, esta é uma evidência que resultará de muitas referências se reportarem a factos e não propriamente opiniões. Ainda assim, constata-se que são mais as opiniões negativas do que as positivas sobre a matéria.

- Modelo Naive Bayes

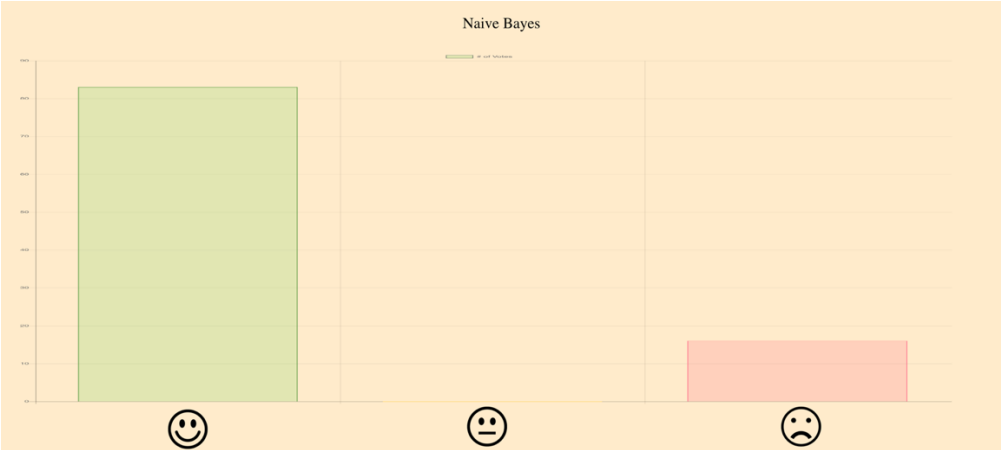


Figura 34: Análise Naive Bayes do tema "corruption".

Através da análise pelo modelo Naive Bayes os resultados obtidos demonstraram que, em relação ao tema *"corruption"*, existe 83.02% de positividade, 0.06% de neutralidade, e 16.92% de negatividade. Este resultado não deixa de causar alguma surpresa. Porém, ele espelhará o facto de se tratar de um modelo classificador treinado, especificamente, num contexto de críticas cinematográficas. Tal pode, assim, ter influenciado o desempenho deste modelo na análise das opiniões recolhidas.

4.9.4 O Tema "Ryanair"

A companhia aérea *Ryanair* tem estado no centro de várias disputas laborais, tendo sido notícia nos vários meios de comunicação social pelos mais variados motivos. Além dos problemas habituais com atrasos de voos, com o tratamento das malas dos passageiros, entre outros, o comportamento por parte da empresa *Ryanair* tem sido alvo de críticas bastante negativas, tanto por parte dos seus próprios funcionários como das pessoas que utilizam os seus serviços. Deste modo, considerámos o tema *"ryanair"* como um bom tema para análise, com a expectativa de obter resultados bem descritivos e correspondentes à atual opinião social sobre esta mesma empresa.

No processo de análise realizado, tal como é apresentado de seguida, o modelo padrão aproxima-se mais da realidade da opinião social relativamente ao assunto "ryanair", analisando grande parte dos 10000 tweets extraídos como negativos. De facto, o resultado, na sua grande parte, resulta, em primeiro lugar, de comentários negativos sobre a empresa, e em segundo lugar de comentários neutros.

- Modelo padrão



Figura 35: Análise padrão do tema "ryanair".

Através da análise realizada pelo modelo padrão observou-se que, em relação ao tema *"ryanair"* os valores obtidos são de 29.47% de positividade, 31.87% de neutralidade e 38.66% de negatividade. Os resultados obtidos expressarão a opinião social existente sobre a empresa. A percentagem atribuída à negatividade é, sem dúvida, um valor que poderá ser explicado com o facto da empresa ter visto o seu nome envolvido em polémicas várias, ao nível da qualidade do serviço que presta e da relação laboral que mantém com os seus colaboradores.

- Modelo Naive Bayes

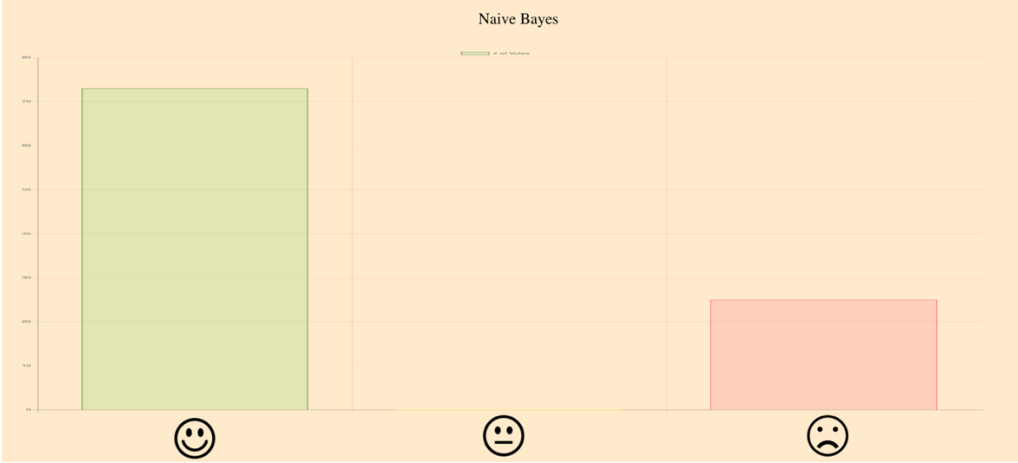


Figura 36: Análise Naive Bayes do tema "ryanair".

Através da análise pelo modelo Naive Bayes notou-se que, em relação ao tema *"ryanair"*, existe 73.78% de positividade, 0.46% de neutralidade, e 25.76% de negatividade. Relativamente a este tema, o modelo Naive Bayes apresentou um resultado aquém do esperado. Com efeito, as opiniões positivas demonstrarão que para muitos dos utilizadores, o custo é um fator decisivo no momento de optar.

4.9.5 O Tema "McDonald's"

A McDonald's é reconhecida como a maior cadeia de restaurantes de *fast food*. Além disso, também é conhecida pelos muitos processos que enfrenta na justiça devido à origem dos seus produtos, aos malefícios que o tipo de produtos comercializados pela empresa provoca no ser humano ou ao excesso de plástico que esta cadeia origina, uma vez que comercializa os seus produtos recorrendo a muitos artigos de plástico, entre muitos outros. Apesar de todos os lucros alcançados, existem vários fatores negativos em torno da empresa, em particular a existência de um fator negativo bastante forte associado com os malefícios que os produtos da cadeia McDonald's trazem para o ser humano.

Após a extração dos 10000 tweets referentes ao tema "mcdonalds" pode-se observar que no modelo padrão, apesar da taxa relativa à positividade ser de 40%, a taxa relativa à neutralidade apenas fica meros pontos percentuais atrás seguida logo da taxa relativa à negatividade. Por outro lado, quando analisados pelo modelo Naive Bayes, os 10000 tweets extraídos apresentam resultados maioritariamente negativos. Neste caso, pode-se considerar que este tema é tendencialmente negativo e, por isso, considera-se que o modelo de Naive Bayes apresenta melhor performance, uma vez que, apesar de não deixar de ter uma taxa percentual positiva alta, possui uma taxa percentual negativa ainda maior.

- Modelo padrão



Figura 37: Análise padrão do tema "mcdonalds".

Através da análise realizada pelo modelo padrão os resultados obtidos demonstraram que, em relação ao tema *"mcdonald's"*, os valores obtidos são de 42.04% de positividade, 34.72% de neutralidade e 23.24% de negatividade. Os resultados desta análise revelam uma divisão clara dos utilizadores, sendo a maioria das opiniões favoráveis. Não deixa de ser assinalável que a percentagem de opiniões neutras é significativa, sendo que as negativas são superiores a metade das positivas. Ou seja, a soma das opiniões neutras e negativas é superior às opiniões positivas, mostrando que existe uma grande heterogeneidade de opiniões.

- Naive Bayes

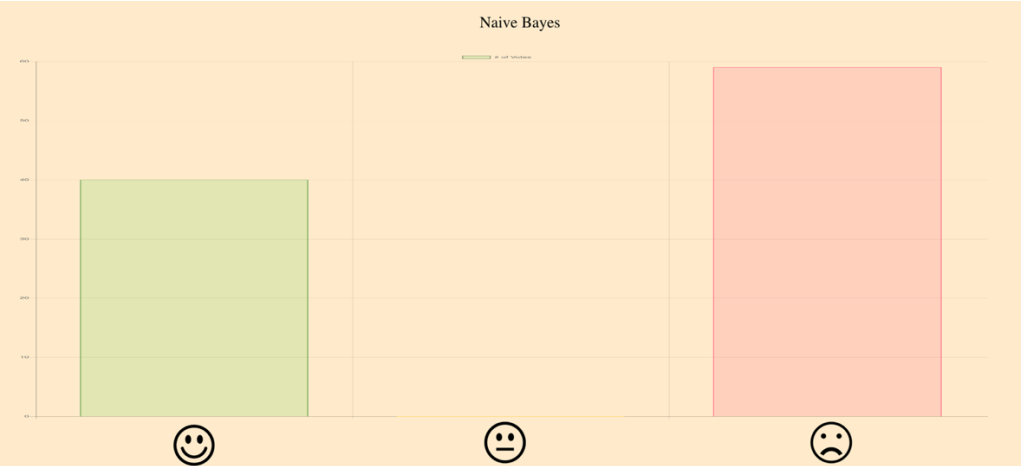


Figura 38: Análise Naive Bayes do tema "mcdonalds".

Através da análise pelo modelo Naive Bayes observou-se que, em relação ao tema *"mcdonalds"*, existe 40.27% de positividade, 0.55% de neutralidade, e 59.18% de negatividade. Na análise realizada por este modelo sobre este mesmo tema, verifica-se a radicalização das opiniões. A vantagem é, claramente, para as opiniões negativas, com quase 20% de supremacia em relação às opiniões consideradas positivas. De salientar a quase inexistência de opiniões neutras em relação a este tema.

4.9.6 Conclusão dos Resultados dos Modelos

Obtidos os resultados dos modelos classificadores em relação aos *tweets* extraídos verifica-se, com alguma evidência, alguma disparidade. Com efeito, alguns dos resultados obtidos, não permitem perceber uma ideia clara dos

utilizadores em relação a um determinado tema, se tivermos em conta a comparação das duas análises entre si. De notar que o modelo Naive Bayes foi treinado com um conjunto de dados de treino relativos a críticas de cinema, o que, por vezes, influencia negativamente o desempenho do modelo num contexto de críticas de outros temas.

Ainda assim, os resultados obtidos pelos modelos não deixam de expressar opiniões que não podem ser ignoradas. Resulta daqui que o modelo padrão apresentou indicadores positivos sobre as análises pretendidas.

5 Conclusões e Trabalho Futuro

5.1 Conclusões

O mundo da tecnologia está constantemente em inegável evolução e, como tal, providencia enormes possibilidades de avanço em todos os campos associados a cada uma das suas áreas de aplicação. A revolução tecnológica obriga a uma constante evolução dos métodos inerentes a tudo o que rodeia o mercado tecnológico e, por isso mesmo, o estudo no campo da mineração de opiniões e de sentimentos, para estabelecimento de índices de satisfação, pode ser visto como contínuo e, na nossa opinião, deve, regularmente, procurar melhorar os processos que se inserem no desenvolvimento de um dado sistema com aplicação numa dada temática área de desenvolvimento tecnológica.

A mineração de opiniões e de sentimentos é, indubitavelmente, uma área de estudo e desenvolvimento crucial, ainda mais quando a situação atual é de uma revolução tecnológica. Em casos como estes, são gerados, diariamente, enormes volumes de dados que contribuem, cada vez mais, para o enriquecimento do conhecimento dos algoritmos que os analisam. A cada momento que passa, verificamos a publicação de fotografias, de novas notícias, *tweets*, ou de estados, localizações e críticas sobre hotéis, sobre restaurantes, ou sobre outra coisa qualquer. Toda esta quantidade de dados constitui um valor bastante significativo para a compreensão de tendências sociais, políticas, de mercado, e de muitas outras mais. Além do mais gera interesse para vários agentes de mercado sejam eles empresas ou consumidores.

Uma empresa pode recolher dados juntos dos consumidores que foram ou são alvo de acções de prestação de serviços ou de aquisição de produtos por parte da mesma, e estes mesmos dados, depois de devidamente pré-

processados, analisados e disponibilizados, podem beneficiar intuitivamente o *feedback* que obrigatoriamente deve existir para a contínua melhoria dos serviços prestados pela empresa e para o contínuo progresso no desenvolvimento de produtos adequadamente concebidos para a satisfação do consumidor final. O suporte de um desenvolvimento sustentado, bem fundamentado e alicerçado de uma empresa deve coexistir com um estudo intensivo das reações que os produtos e serviços prestados pela empresa junto dos clientes podem gerar. Por outro lado, é significativamente interessante conhecer as opiniões de quem já utilizou determinado produto ou serviço, para que assim, o próximo indivíduo a (pretender) adquirir um produto ou serviço da mesma gama possa fazer uso desta experiência partilhada e obter conhecimento de modo a poder fazer uma escolha sustentada sobre o produto ou serviço que vai, efetivamente, adquirir.

Atualmente a maior parte das pessoas que deseja realizar uma viagem costuma executar uma pesquisa para encontrar um hotel de acordo com as características que pretende, e, além da lista de hotéis, que se assemelham com o conjunto de elementos que definiu como as características pretendidas tem também detalhadamente, e geralmente, uma classificação do hotel baseada nas críticas que outros utilizadores que já frequentaram o hotel realizaram.

Tal é benéfico para quem nunca esteve no hotel, uma vez que pode ler todas as críticas e decidir se realmente é um hotel merecedor da escolha que irá fazer. O mesmo pode afirmar-se para o caso de se pretender adquirir um automóvel. Neste caso, o utilizador procura marcas e modelos de automóveis para os quais não sejam conhecidos problemas frequentes e para os quais a relação preço/qualidade seja razoável. Assim, pode afirmar-se que, hoje em dia, esse tipo de informações, usualmente disponibilizada através de dados textuais, contém um valor imensamente elevado para os utilizadores. Não obstante a complexidade do processamento deste tipo de dados textuais e sua consequente análise, estes são uma categoria de dados que permite, de forma mais específica, tomar conhecimento sobre as opiniões e sentimentos dos consumidores.

(Zhan, Loh e Liu, 2009) enunciaram a existência de múltiplos métodos de mineração de opiniões relativamente ao processamento de críticas de produtos. A conclusão realça que a grande maioria destes sistemas se concentra exclusivamente na contabilização de comentários positivos e negativos por partes dos autores das críticas. Para dar resposta a esta lacuna de disponibilização dos dados analisados, os índices de satisfação são criados tendo em vista a forma intuitiva com que são consultados, e a facilidade com que tornam possível a retirada de ilações por parte de quem se encontra a fazer uso deles mesmos. Deste modo, a aplicação implementada no âmbito deste trabalho, foi arquitetada e idealizada como uma aplicação que possui como matéria-prima as opiniões e os sentimentos. A escolha para fontes de dados recaiu sobre o *Twitter* uma vez que este é uma fonte rica em dados textuais.

Integrada num conjunto de objetivos pré-definidos e com um propósito específico, a aplicação providencia índices de satisfação relativamente aos temas inseridos bem como a justificação para tais índices demonstrando e explicitando a análise que foi realizada e com base em que parâmetros avaliativos. Os benefícios trazidos pela aplicação para uma empresa ou para um indivíduo são significativos, uma vez que as perspetivas providenciadas ao utilizador, seja ele de que tipo anteriormente especificado for, são imensas e permitem a avaliação de certos fatores de mercado. Além de uma análise instantânea do tema inserido, é possível ainda realizar uma análise cronológica o que, por sua vez, permite obter uma análise específica e previsão do progresso que o tema inserido pode vir a realizar no mercado.

Na secção 4.9 desta dissertação foi realizada a comparação de resultados para uma amostra razoavelmente significativa de 10000 *tweets*, verificando-se que dos 5 (cinco) temas analisados, os resultados obtidos por cada um dos modelos em análise se assemelharam razoavelmente.

5.2 Trabalho Futuro

A área da mineração de opiniões e de sentimentos é uma área em contínua expansão. Esta área permite por sua vez o progresso e a melhoria contínua do modo como ocorre a disponibilização de índices de satisfação e outros indicadores da expressividade e opiniões que são partilhadas nas mais diversas plataformas *Web*. Através deste processo, podemos recorrer, por exemplo, aos dados recolhidos pelas empresas aquando da prestação de serviços junto de clientes sobre o suporte prestado, por exemplo, algo que é sempre um objetivo contínuo de toda e qualquer empresa. As empresas fazem uso destes dados, para, ao longo do tempo, melhorarem o serviço prestado e com isto, poder crescer no mercado em termos, sobretudo, de aceitação daquilo que oferecem ao seu público-alvo. Um bom cliente, ao qual é providenciado suporte de qualidade por parte de uma empresa, torna-se num cliente que propaga a outro potencial cliente a qualidade existente dessa empresa, beneficiando-a posteriormente nos seus processos de venda de produtos e de serviços a esse potencial novo cliente.

Atualmente, em várias plataformas sociais como é o caso do *Facebook* e do *Instagram*, os perfis dos utilizadores são gradualmente construídos, para que, deste modo, possam proporcionar aos seus clientes as melhores sugestões de produtos a adquirir. Muitas das vezes, e inseridos nas categorias de produtos pesquisados recentemente pelo utilizador, as plataformas sugerem novas sugestões de compra. Aliada à mineração de opiniões e de sentimentos esta pode ser uma ferramenta para revolucionar todo e qualquer mercado. As empresas podem, deste modo, compreender e arquitetar perfis de utilizadores, redirecionando com maior eficácia os seus clientes-alvo para os produtos que estes mais desejam ou mais se sentem inclinados para adquirir.

Indubitavelmente, a revolução tecnológica que tem vindo a acontecer fez com que as pessoas alterassem a forma como expressam os seus pontos de vista e as suas opiniões. Tal pode, sem dúvida, ser aproveitado pelas empresas para, gradualmente, recolher e analisar as opiniões e sentimentos partilhados pelas pessoas, para que sejam usadas a favor de um aumento do lucro obtido pela empresa com as suas vendas direcionadas cada vez melhor para o seu público-alvo, baseando-se em regras de análise fundamentadas e providenciadas por sistemas de extração e de análise de opiniões e de sentimentos que permitem disponibilizar um conjunto de índices de satisfação e de análise aprofundadas sobre os dados recolhidos.

Hoje, é possível observar que qualquer pessoa partilha críticas dos produtos que comprou a certa empresa, sendo que esta define, à vista de todos, a imagem de uma marca, de uma empresa. Assim, é fundamental que as empresas saibam aplicar o seu melhor conhecimento a cada cliente em específico de modo a que cada um saiba referir aquilo que acha que é o melhor da empresa. A presença em fóruns, em grupos de discussão ou em blogs de comentários opinativos, sobre produtos e serviços faz com que seja possível criar um índice de satisfação como aquele que foi calculado neste trabalho de dissertação, fazendo com que a opinião social funcione como um mercado de índices de classificação com base nas opiniões que são partilhadas sobre cada um dos produtos e serviços comercializados pelas empresas.

A aplicação implementada no âmbito deste trabalho, integrada num conjunto de objetivos pré-definidos, serve o propósito de providenciar índices de satisfação relativamente aos temas inseridos bem como a justificação para tais índices, demonstrando e explicitando a análise que foi realizada e que parâmetros avaliativos foram tomados em consideração. Um dos fatores que melhoraria, indubitavelmente, a performance do sistema desenvolvido no âmbito desta dissertação seria a implementação de processos de treino adequados para cada um dos modelos utilizados. Além deste facto, pode ainda notar-se que, na análise cronológica realizada pelo sistema, foi objetivo inicial implementar esta

característica tendo por base uma relação ano/mês. Esta perspetiva permite uma visualização de enorme utilidade, uma vez que permite a análise contínua de um tema ao longo do tempo, providenciando elementos justificadores para uma melhor tomada de decisão a longo prazo.

Outro dos trabalhos futuros que podemos ter em consideração é a implementação de um modelo customizado para a língua portuguesa, no qual possam ser definidos conjuntos de treino para treinar um modelo especificamente idealizado e preparado para lidar com a complexidade da língua portuguesa. Dessa forma, seria possível fazer a análise de assuntos com maior especificidade e inerentes a Portugal. A implementação do sistema descrito, incrementaria significativamente a capacidade das empresas portuguesas em darem uma resposta mais efetiva a um mercado cada vez mais competitivo a nível mundial.

6 Bibliografia

Abbasi, A. (2010) 'AI and opinion mining, part 2', in *IEEE Intelligent Systems*, pp. 72–79. doi: 10.1109/MIS.2010.94.

Anjaria, M. and Guddeti, R. M. R. (2014) 'Influence factor based opinion mining of Twitter data using supervised learning', *Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference on*. doi: 10.1109/COMSNETS.2014.6734907.

Archak, N., Ghose, A. and Ipeirotis, P. G. (2011) 'Deriving the Pricing Power of Product Features by Mining Consumer Reviews', *Management Science*, 57(8), pp. 1485–1509. doi: 10.1287/mnsc.1110.1370.

Balazs, J. A. and Velásquez, J. D. (2016) 'Opinion Mining and Information Fusion: A survey', *Information Fusion*, 27, pp. 95–110. doi: 10.1016/j.inffus.2015.06.002.
Bhadane, C., Dalal, H. and Doshi, H. (2015) 'Sentiment Analysis: Measuring Opinions', *Procedia Computer Science*, 45, pp. 808–814. doi: 10.1016/j.procs.2015.03.159.

Bhuiyan, T., Xu, Y. and Jøsang, A. (2009) 'QUT Digital Repository : State-of-the-Art Review on Opinion Mining from Online Customers ' Feedback', *The 9th Asia-Pacific Complex Systems Conference*, (November), pp. 4–7. doi: 10.1.1.155.2842.

Binali, H., Potdar, V. and Wu, C. (2009) 'A state of the art opinion mining and its application domains', in *2009 IEEE International Conference on Industrial Technology*, pp. 1–6. doi: 10.1109/ICIT.2009.4939640.

Breault, J. L., Goodall, C. R. and Fos, P. J. (2002) 'Data mining a diabetic data warehouse', *Artificial Intelligence in Medicine*. doi: 10.1016/S0933-3657(02)00051-9.

Breck, E., Choi, Y. and Cardie, C. (2007) 'Identifying expressions of opinion in context', in *IJCAI International Joint Conference on Artificial Intelligence*. doi: 10.1016/j.jad.2005.02.015.

Cambria, E. *et al.* (2013) 'New Avenues in Opinion Mining and Sentiment Analysis', *IEEE Intelligent Systems*. doi: 10.1109/MIS.2013.30.

Carvalho, A. and Belo, O. (2016) 'Predicting waiting time in customer queuing systems', in *2016 IEEE International Conference on Knowledge Engineering and Applications, ICKEA 2016*. doi: 10.1109/ICKEA.2016.7803010.

Chen, H. (2011) 'Dark Web: Exploring and Mining the Dark Side of the Web', in *2011 European Intelligence and Security Informatics Conference*, pp. 1–2. doi: 10.1109/EISIC.2011.78.

Chen, H. (2012) 'Sentiment Analysis', *Dark Web*, 30, pp. 171–201. doi: 10.1007/978-1-4614-1557-2_10.

Chen, H. and Zimbra, D. (2010) 'AI and opinion mining', in *IEEE Intelligent Systems*, pp. 74–76. doi: 10.1109/MIS.2010.75.

Chesley, P. *et al.* (2006) 'Using verbs and adjectives to automatically classify blog sentiment', *Training*. doi: 10.1016/0020-0271(73)90092-2.

Church, K. W. and Hanks, P. (1989) 'Word association norms, mutual information, and lexicography', in *Proceedings of the 27th annual meeting on Association for*

Computational Linguistics -. doi: 10.3115/981623.981633.

Das, S. R. and Chen, M. Y. (2007) 'Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web', *Management Science*. doi: 10.1287/mnsc.1070.0704.

Dave, K. *et al.* (2003) 'Mining the peanut gallery: Opinion extraction and semantic classification of product reviews', *Proceedings of the 12th international conference on World Wide Web*. doi: 10.1145/775152.775226.

Dellarocas, C., Zhang, X. (Michael) and Awad, N. F. (2007) 'Exploring the value of online product reviews in forecasting sales: The case of motion pictures', *Journal of Interactive Marketing*. doi: 10.1002/dir.20087.

Devika, M. D., Sunitha, C. and Ganesh, A. (2016) 'Sentiment Analysis: A Comparative Study on Different Approaches', *Procedia Computer Science*, 87, pp. 44–49. doi: 10.1016/j.procs.2016.05.124.

Devitt, A. and Ahmad, K. (2007) 'Sentiment polarity identification in financial news: a cohesion-based approach', *Proceedings of Annual Meeting of the Association of Computational Linguistics*. doi: 10.1.1.143.7157.

Dey, L. and Haque, S. M. (2009) 'Opinion mining from noisy text data', in *International Journal on Document Analysis and Recognition*, pp. 205–226. doi: 10.1007/s10032-009-0090-z.

Dokoohaki, N. *et al.* (2015) 'Predicting Swedish Elections with Twitter', in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15*, pp. 1269–1276. doi: 10.1145/2808797.2808915.

Eguchi, K. and Lavrenko, V. (2006) 'Sentiment retrieval using generative models', in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing - EMNLP '06*. doi: 10.3115/1610075.1610124.

Fano, R. M. and Hawkins, D. (1961) 'Transmission of Information: A Statistical Theory of Communications', *American Journal of Physics*. doi: 10.1119/1.1937609.

Fornell, C. *et al.* (1996) 'The American Customer Satisfaction Index: Nature, Purpose, and Findings', *Journal of Marketing*. doi: 10.2307/1251898.

Harré, R. and Parrott, W. G. (1996) *The emotions*, *Studies in emotion and social interaction.*

Hatzivassiloglou, V. and McKeown, K. R. (1997) 'Predicting the semantic orientation of adjectives', in *Proceedings of the 35th annual meeting on Association for Computational Linguistics* -. doi: 10.3115/976909.979640.

Hatzivassiloglou, V. and Wiebe, J. M. (2000) 'Effects of adjective orientation and gradability on sentence subjectivity', in *Proceedings of the 18th conference on Computational linguistics* -. doi: 10.3115/990820.990864.

Hong, Y. and Skienay, S. (2010) 'The Wisdom of Bookies? Sentiment Analysis vs. the NFL Point Spread', *Association for the Advancement of Artificial Intelligence wwwaaaiorg*, pp. 251–254.

Howe, N. and Cardie, C. (1997) 'Improving Minority Class Prediction Using Case-Specific FeatureWeights', in *Proceedings of the Fourteenth International Confernece on Machine Learning*.

Hu, M. and Liu, B. (2004a) 'Mining and summarizing customer reviews', in *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*. New York, New York, USA: ACM Press, p. 168. doi: 10.1145/1014052.1014073.

Hu, M. and Liu, B. (2004b) 'Mining Opinion Features in Customer Reviews', *19th national conference on Artifical intelligence*, pp. 755–760. doi: 10.1145/1014052.1014073.

Jindal, N. and Liu, B. (2006) 'Identifying comparative sentences in text documents', in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*. doi: 10.1145/1148170.1148215.

Jindal, N. and Liu, B. (2008) 'Opinion spam and analysis', in *Proceedings of the international conference on Web search and web data mining - WSDM '08*, p. 219. doi: 10.1145/1341531.1341560.

Jindal, N., Liu, B. and Bing, L. (2006) 'Mining comparative sentences and relations', *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*. doi: 10.1107/S0108270189000326.

Khan, K. *et al.* (2009) 'Mining opinion from text documents: A survey', in *2009 3rd IEEE International Conference on Digital Ecosystems and Technologies, DEST '09*, pp. 217–222. doi: 10.1109/DEST.2009.5276756.

Knight, K. and Hatzivassiloglou, V. (1995) 'Two-level, many-paths generation', *Proceedings of the 33rd annual meeting on Association for Computational Linguistics* -. doi: 10.3115/981658.981692.

Landauer, T. K. and Dumais, S. T. (1997) 'A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge', *Psychological Review*. doi: 10.1037/0033-295X.104.2.211.

Landauer, T. K., Folt, P. W. and Laham, D. (1998) 'An introduction to latent semantic analysis', *Discourse processes*. doi: 10.1080/01638539809545028.

Lee, D., Jeong, O.-R. and Lee, S. (2008) 'Opinion mining of customer feedback data on the web', in *Proceedings of the 2nd international conference on Ubiquitous information management and communication - ICUIMC '08*, p. 230. doi: 10.1145/1352793.1352842.

Liu, B. (2010) 'Sentiment Analysis and Subjectivity', *Handbook of Natural Language Processing*, (1), pp. 1–38. doi: 10.1145/1772690.1772756.

Liu, B. (2012) 'Sentiment Analysis and Opinion Mining', *Synthesis Lectures on Human Language Technologies*, 5(1), pp. 1–167. doi: 10.2200/S00416ED1V01Y201204HLT016.

Liu, B. and Zhang, L. (2012) 'A survey of opinion mining and sentiment analysis', in *Mining Text Data*, pp. 415–463. doi: 10.1007/978-1-4614-3223-4_13.

Liu, X., Zhai, K. and Pedrycz, W. (2012) 'An improved association rules mining method', *Expert Systems with Applications*. doi: 10.1016/j.eswa.2011.08.018.

McGlohon, M., Glance, N. and Reiter, Z. (2010) 'Star Quality : Aggregating Reviews to Rank Products and Merchants', *Wall Street Journal*, pp. 114–121. Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewPDFInterstitial

/1507/1844.

McKeown, K. R. *et al.* (1999) 'Towards Multidocument Summarization by Reformulation: Progress and Prospects', *Proceedings of the National Conference on Artificial intelligence*.

Medhat, W., Hassan, A. and Korashy, H. (2014) 'Sentiment analysis algorithms and applications: A survey', *Ain Shams Engineering Journal*, 5(4), pp. 1093–1113. doi: 10.1016/j.asej.2014.04.011.

Metsis, V., Androutsopoulos, I. and Paliouras, G. (2006) 'Spam filtering with naive bayes-which naive bayes?', *Ceas*. doi: 10.1.1.61.5542.

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (1983) *Machine learning: An artificial intelligence approach.*, *Machine Learning: An Artificial Intelligence Approach*. doi: 10.1109/MWSYM.2007.380062.

Mishra, M., Khandelwal, D. and Mishra, V. K. (2016) 'Simple Learner Based Polarity Prediction on Twitter Data', *i-Manager's Journal on Computer Science*.

Morinaga, S. *et al.* (2002) 'Mining product reputations on the Web', in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*. doi: 10.1145/775094.775098.

Najmi, E. *et al.* (2015) 'CAPRA: a comprehensive approach to product ranking using customer reviews', *Computing*, 97(8), pp. 843–867. doi: 10.1007/s00607-015-0439-8.

Nasukawa, T. and Yi, J. (2003) 'Sentiment Analysis : Capturing Favorability Using Natural Language Processing', *Proceedings of the 2nd international conference*

on Knowledge capture. doi: 10.1145/945645.945658.

Nielsen, F. Å. (2011) 'A new ANEW: Evaluation of a word list for sentiment analysis in microblogs', *Knowledge-Based Systems*, 89, pp. 14–46. doi: 10.1016/j.knosys.2015.06.015.

Novalić, A. (2013) *Introduction to tweepy, Twitter for Python - Python Central*, *PythonCentral*.

O'Connor, B. *et al.* (2010) 'From tweets to polls: Linking text sentiment to public opinion time series', *From tweets to polls: Linking text sentiment to public opinion time series*, (May), pp. 122–129. doi: citeulike-article-id:7044833.

Ortigosa, A., Martín, J. M. and Carro, R. M. (2014) 'Sentiment analysis in Facebook and its application to e-learning', *Computers in Human Behavior*. doi: 10.1016/j.chb.2013.05.024.

Pang, B. and Lee, L. (2008) 'Opinion Mining and Sentiment Analysis', *Found. Trends Inf. Retr.*, 2, pp. 1–135. doi: 10.1561/1500000011.

Pang, B., Lee, L. and Vaithyanathan, S. (2002) 'Thumbs up? Sentiment Classification using Machine Learning Techniques', in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - EMNLP '02*. doi: 10.3115/1118693.1118704.

Patodkar, V. N. and I.R, S. (2016) 'Twitter as a Corpus for Sentiment Analysis and Opinion Mining', *IJARCCE*, 5(12), pp. 320–322. doi: 10.17148/IJARCCE.2016.51274.

Rapaport, W. J. (1984) 'Quasi-indexical reference in propositional semantic

networks', *… of the 10th International Conference on …*.

Raut, S. N. (2016) 'User Analytics-Based System Using Social Media Stream Data', *International Journal of Engineering Science and Computing*.

Ravi, K. and Ravi, V. (2015) 'A survey on opinion mining and sentiment analysis: Tasks, approaches and applications', *Knowledge-Based Systems*, 89, pp. 14–46. doi: 10.1016/j.knosys.2015.06.015.

Russell, M. a (2011) *Mining Twitter*, *Russell The Journal Of The Bertrand Russell Archives*. doi: 10.3758/s13428-011-0116-6.

Santorini, B. (1990) 'Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision)', *University of Pennsylvania 3rd Revision 2nd Printing*. doi: 10.1017/CBO9781107415324.004.

Smadja, F., McKeown, K. R. and Hatzivassiloglou, V. (1996) 'Translating collocations for bilingual lexicons: A statistical approach', *Computational linguistics*.

Soler, J. M., Cuartero, F. and Roblizo, M. (2012) 'Twitter as a tool for predicting elections results', in *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012*, pp. 1194–1200. doi: 10.1109/ASONAM.2012.206.

Somprasertsri, G. (2010) 'Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization', *Computer*, 16(6), pp. 938–955. doi: 10.3217/jucs-016-06-0938.

Sonai Muthu Anbananthen, Kalaiarasi and Mohamed H. Elyasir, A. (2013)

'Evolution of Opinion Mining', *Australian Journal of Basic and Applied Sciences*.

Spertus, E. (1997) 'Smokey: Automatic Recognition of Hostile Messages', in *AAAI'97/IAAI'97 Proceedings of the fourteenth national conference on artificial intelligence and ninth conference on Innovative applications of artificial intelligence*, pp. 1058–1065.

Stoyanov, V. *et al.* (2006) 'Evaluating an opinion annotation scheme using a new multi-perspective question and answer corpus', in *Computing Attitude and Affect in Text: Theory and Applications*. doi: 10.1007/1-4020-4102-0_8.

Taboada, M. *et al.* (2011) 'Lexicon-Based Methods for Sentiment Analysis', *Computational Linguistics*, 37(2), pp. 267–307. doi: 10.1162/COLI_a_00049.

Tatemura, J. (2000) 'Virtual reviewers for collaborative exploration of movie reviews', *Proceedings of the 5th international conference on …*. doi: 10.1145/325737.325870.

Terveen, L. *et al.* (1997) 'PHOAKS: a system for sharing recommendations', *Communications of the ACM*. doi: 10.1145/245108.245122.

Tiwari, A., Turner, C. J. and Majeed, B. (2008) 'A review of business process mining: state-of-the-art and future trends', *Business Process Management Journal*, 14(1), pp. 5–22. doi: 10.1108/14637150810849373.

Tumasjan, A. *et al.* (2010) 'Predicting elections with Twitter: What 140 characters reveal about political sentiment', *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pp. 178–185. doi: 10.1074/jbc.M501708200.

Turney, P. D. (2001) 'Mining the Web for synonyms: PMI-IR versus LSA on TOEFL', *Proceedings of the 12th European Conference on Machine Learning (ECML-2001), Freiburg, Germany*. doi: 10.1007/3-540-44795-4_42.

Turney, P. D. (2001) 'Thumbs up or thumbs down?', in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. Morristown, NJ, USA: Association for Computational Linguistics, p. 417. doi: 10.3115/1073083.1073153.

Valarmathi, B. and Palanisamy, V. (2011) 'Opinion Mining Classification Using Key Word Summarization Based on Singular Value Decomposition', *International Journal on Computer Science and Engineering (IJCSE)*. doi: 10.1007/s00383-009-2426-7.

Wicklund, D. A., Palermo, D. S. and Jenkins, J. J. (1964) 'The effects of associative strength and response hierarchy on paired-associate learning', *Journal of Verbal Learning and Verbal Behavior*. doi: 10.1016/S0022-5371(64)80011-6.

Wiebe, J. M. (1990) 'Identifying subjective characters in narrative', *Proceedings of the International Conference on Computational Linguistics (COLING)*. doi: 10.3115/982023.982039.

Wiebe, J. M. (2000) 'Learning subjective adjectives from corpora', *Proceedings of the National Conference on Artificial Intelligence*. doi: http://portal.acm.org/citation.cfm?id=721121&dl=ACM&coll=&CFID=15151515&CFTOKEN=6184618.

Wiebe, J. M., Bruce, R. F. and O'Hara, T. P. (1999) 'Development and use of a gold-standard data set for subjectivity classifications', in *Proceedings of the 37th*

annual meeting of the Association for Computational Linguistics on Computational Linguistics -. doi: 10.3115/1034678.1034721.

Wilks, Y. and Bien, J. (1983) 'Beliefs, points of view, and multiple environments', *Cognitive Science*. doi: 10.1016/S0364-0213(83)80007-X.

Wu, Y. *et al.* (2013) 'OpinionSeer: interactive visualization of hotel customer feedback.', *IEEE transactions on visualization and computer graphics*, 16(6), pp. 1109–18. doi: 10.1109/TVCG.2010.183.

Xia, Y. Q. *et al.* (2007) 'The unified collocation framework for opinion mining', in *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, ICMLC 2007*, pp. 844–850. doi: 10.1109/ICMLC.2007.4370260.

Zhan, J., Loh, H. T. and Liu, Y. (2009) 'Gather customer concerns from online product reviews - A text summarization approach', *Expert Systems with Applications*. doi: 10.1016/j.eswa.2007.12.039.

Zhang, Y. and Zhu, W. (2013) 'Extracting Implicit Features in Online Customer Reviews for Opinion Mining', *Proceedings of the 22Nd International Conference on World Wide Web Companion*, pp. 103–104. doi: 10.1145/2487788.2487835.