Departamento de Ciências e Tecnologias de Informação

Análise de Sentimentos na Classificação de Comentários *Online* Aplicando Técnicas de *Text Mining*

Águeda Cabral Moreno

Dissertação submetida como requisito parcial para obtenção do grau de

Mestre em Sistemas Integrados de Apoio à Decisão

Orientador:
Doutor Paulo Rita – Professor Catedrático
Departamento de Marketing, Operações e Gestão Geral
ISCTE-IUL Business School

Coorientador:
Doutor João Guerreiro - Professor Auxiliar
Departamento de Marketing, Operações e Gestão Geral
ISCTE-IUL

outubro, 2015

“Public sentiment is everything.
With public sentiment nothing can fail.
Without it, nothing can succeed.”
Abraham Lincoln

Agradecimentos

Seguindo as pegadas da pessoa mais forte e corajosa que conheço, tento todos os dias fazer os possíveis para conseguir alcançar os meus objetivos e retribuir todo o carinho, apoio e confiança depositada em mim. Agradeço a minha maravilhosa, batalhadora, paciente e incansável mãe por nunca duvidar das escolhas que fiz e por sempre estar do meu lado em todos os momentos. Ela é, e vai continuar sendo, o meu pilar, o meu porto seguro que, apesar da distância física, sinto-a sempre presente.

Agradeço todo o apoio dado pelo meu orientador e coorientador Prof. Dr. Paulo Rita e Prof. Dr. João Guerreiro, respetivamente, por terem aceitado o meu projeto e por apresentarem sempre disponibilidade para esclarecer as minhas dúvidas e por serem incansáveis e rápidos nas respostas. Isso ajudou-me bastante em todas as fases do projeto e aprendi muito com eles. Obrigada por partilharem o vosso conhecimento comigo e os demais colegas do ISCTE-IUL.

Agradeço também, ao meu pai, meus irmãos, familiares e todos os meus amigos, por todo o apoio e compreensão, por muitas vezes não poder estar presente nos convívios sociais e por ter recusado muitos convites e estado ausente durante todo este período.

A todos os meus amigos e colegas de trabalho da Morphis Tech que direta ou indiretamente apresentaram o seu apoio e estiveram sempre a tentarem motivar-me, em especial ao Fernando Ramalho, Cristina Teles, Márcia Morgado, que foram compreensivos relativamente às minhas ausências apesar de, ainda, ter muito pouco tempo na equipa.

Resumo

O crescimento dos *sociais media* proporcionou, nos últimos anos, um aumento significativo de comentários *online* que se refletem nas decisões de compra. Os comentários ajudam, por um lado, as empresas a recolher informações quanto à perceção dos consumidores em relação aos seus bens e serviços. Por outro lado, ajudam e influenciam os consumidores a centrarem a sua atenção nas recomendações que poderão estar mais alinhadas em satisfazer as suas necessidades, filtrando à partida uma grande quantidade de informação que poderá não responder a esses requisitos.

O presente projeto tem como objetivo dar resposta a esta problemática através do estudo da plataforma Yelp. Para tal, foram extraídos 14.000 comentários, relacionados com diferentes produtos turísticos, com os respetivos votos (*useful)*. Sobre estes foram aplicadas técnicas de *text mining* de modo a encontrar os principais sentimentos (positivos, neutros e negativos), tópicos e termos de cada comentário, que permitem explicar a sua utilidade.

Durante a investigação, seguindo a metodologia CRISP-DM, organizou-se os comentários em tópicos, construiu-se o *wordcloud* com os termos mais utilizados pelos consumidores, procedeu-se à análise de sentimentos dos comentários, das entidades e dos tópicos correspondentes e, por último, construíram-se quatro modelos preditivos, calculando os erros de treino e teste.

Os resultados obtidos mostram que o modelo Regressão Logística é o melhor dos modelos construídos, onde os termos: *chair, valley, neighborhood* e *place food* são os mais importantes para explicar a utilidade dos comentários. Agruparam-se ainda os comentários em 20 tópicos onde o tópico "Buffet" revelou ser o mais útil e com sentimento positivo.

Palavras-chave: *Text Mining*, Comentários *Online*, Análise de Sentimentos, *Data Mining*, Processamento de Linguagem Natural, *Topic Model.*

Classificação ACM: H.2.8 *Database Applications – Data Mining,* H.4.2 *Types of Systems – Decision Support*

Abstract

The growth of social media lead, in the past few years, to a significant increase of the online reviews that reflect buying decisions. These reviews, on one hand facilitate companies to acquire information regarding the perception of consumers, but on the other hand help and influence consumers to center their attention on the reviews that are better suited for their needs, thus filtering a huge amount of irrelevant information to meet their requirements.

This project aims at addressing these issues and give some useful answers by using the Yelp platform. This study involves the extraction of 14.000 reviews, related to different tourism products, with the respective votes (useful). Text mining techniques were applied in order to identify and extract the main subjective sentiments (positives, negatives and neutral), topics and terms behind each review which then enabled us to understand the usefulness of the reviews.

Throughout this study and by using the CRISP-DM methodology, the researcher organized the reviews by topics, has built a word cloud of the most used terms, performed the sentiment analysis on the reviews, the entities and related topics and finally built the predictive models, by measuring the train and test set.

The results show that the logistic regression model is the best predictive model, where the terms*:* chair*,* valley*,* neighborhood and place food are the most important to explain the usefulness of comments. Still, it was possible to group the comments on 20 topics where the topic “Buffet” proved to be the most useful and positive sentiment.

Keywords: Text Mining, Online Reviews, Sentiment Analysis, Data Mining, Natural Processing Language, Topic Model

ACM Classification: H.2.8 Database Applications – Data Mining, H.4.2 Types of Systems – Decision Support

Índice

Índice de Tabelas

Índice de Figuras

Lista de Siglas e Acrónimos

ACC	*Classification Accuracy Rate*
BOW	*Bag-of-words*
CRISP-DM	*Cross Industry Standard Process for Data Mining*
CRM	*Customer Relationship Management*
CTM	*Correlated Topic Model*
DTM	*Document-by-term matrix*
LDA	*Latent Dirichlet Allocation*
NLP	*Natural Language Process*
PLN	Processamento de Linguagem Natural
POS	*Part-of-speech*
ROC	*Receiver Operating Characteristic*
SVM	*Support Vector Machines*
TF-IDF	*Term Frequency Inverse Document Frequency*
WOM	*word-of-mouth*

1 Introdução

Há cada vez mais consumidores com acesso à *Internet* e que vêm ganhando interesse e dependência pelos serviços disponibilizados na *web.* Esta interação no mundo virtual traduziu-se num grande volume de dados gerados diariamente, quer seja através das redes sociais, *blogs*, fóruns (Pang & Lee, 2008) ou plataformas de recomendações. Na origem deste fenómeno está a *Web 2.0* que facilitou o acesso à informação *online* e a interação entre os utilizadores, criando assim o conceito de *social media* que vem crescendo nos últimos anos e proporcionando um aumento exponencial de informação não estruturada em formato eletrónico. Essa informação está cada vez mais acessível ao consumidor para o ajudar e influenciar na sua tomada de decisão (Greenberg, 2010; Guernsey, 2000). Os comentários *online* ganharam assim uma importância fundamental na decisão, quer dos consumidores quer das empresas. Do lado das empresas, porque estas podem acompanhar como os seus bens e serviços estão a ser apreciados pelo seu consumidor, o que se poderá traduzir-se no crescimento reputacional da marca. Do lado dos consumidores, a recomendação de bens e serviços ajuda-os a centrarem a sua atenção naqueles que poderão estar mais alinhados com a satisfação das suas necessidades, filtrando à partida uma grande quantidade de bens e serviços que poderão não preencher esses requisitos.

Criar e executar este filtro não é fácil, pois no meio de tantos dados surge sempre o problema de confiabilidade, fidedignidade e utilidade destas informações (Hajas *et al.*, 2014) que podem muitas vezes escapar aos olhos mais atentos. Para isso, é preciso que as informações disponíveis nos *social media* estejam atualizadas e sejam, de facto, úteis para essa tomada de decisão. Empresas como a Yelp disponibilizaram plataformas de recomendações que permitem aos seus utilizadores publicarem as suas experiências. No entanto, verifica-se que nem sempre as recomendações são vistas pelos consumidores, como sendo úteis e confiáveis. Os comentários publicados na Yelp podem ser classificados como sendo *Useful, Funny* ou *Cool,* mediante o número de votos que este recebe dos diferentes consumidores.

Litvin *et.al.* (2008) apresentaram um estudo sobre a influência interpessoal e da *word-of-mouth* (WOM) na decisão de compra do consumidor. Eles salientaram a importância destas influências na indústria do turismo devido ao facto do produto ser intangível e, por isso, difícil de ser avaliado antes do consumo. Como base do seu trabalho encontra-se o desenvolvimento de um modelo conceptual WOM para ajudar a perceber influências dos comentários *online* sobre a decisão de compra do consumidor. E como trabalho futuro apresentaram a seguinte questão: "Na ausência do contacto face-a-face com quem publica os comentários, quais são os critérios utilizados para determinar a confiabilidade das influências dos *social media*?" É com base nesta questão que surge o desafio de encontrar os principais sentimentos (positivos, neutros e negativos) e termos por detrás de cada comentário capaz de explicar a sua utilidade com base no voto "useful" que recebem dos consumidores.

A metodologia a seguir na concretização dos objetivo da presente investigação é CRISP-DM, em que será extraída uma amostra aleatória constituída por 14.000 comentários relacionados com vários negócios ligados à indústria turística tais com restaurantes, bares, hotéis, casinos, centros

comerciais, eventos variados, ou seja, tudo o que um turista precisa quando vai visitar um lugar diferente. Esta extração será feita no *software* R, onde também serão realizadas as primeiras transformações dos dados e as primeiras análises descritivas dos documentos, como a construção da *wordcloud* com os termos mais frequentes da amostra. Proceder-se-á à extração das entidades através do *part-of-speech* e a construção de tópicos aplicando o algoritmo *correlated topic model* (CTM) (Blei & Lafferty, 2009). Este permitirá analisar o comportamento e eficácia do CTM aplicado aos comentários *online*. Depois de construídos os tópicos, será elaborado o *profiling* a fim de encontrar a melhor designação para cada um deles. Tendo isto, o próximo passo é a realização da análise de sentimentos no Microsoft Excel com base na aplicação do *plug-in* disponibilizado pelo *software* Semantria. Neste *software* será questionado qual o sentimento presente em cada um dos comentários, bem como os sentimentos das respetivas entidades e tópicos, classificando-os como sendo positivos, negativos ou neutros. Por fim, será construído o modelo preditivo com base nas funcionalidades do *software* IBM SPSS Modeler para tentar encontrar padrões nos comentários passíveis de explicar a sua utilidade.

No geral, os principais objetivos do presente projeto são:

- Agrupar os comentários em diferentes tópicos de acordo com o assunto em que estiverem mais correlacionados;
- Aplicar a análise de sentimentos e descobrir quais os sentimentos por detrás de cada um dos comentários;
- Construir um modelo de classificação capaz de prever os padrões passíveis de explicar a utilidade dos comentários com base no voto “useful”;
- Responder à questão colocada por Litvin *et.al.* (2008) em relação ao critério utilizado pelo consumidor para determinar a confiança de uma influência encontrada nos *social media*;
- Contribuir com mais uma investigação de análise de sentimento para a área do turismo.

Como foi referido anteriormente, a Yelp usa um *software* automático para recomendar os comentários considerados úteis pelos consumidores. O projeto pretende identificar os comentários através das técnicas de análise de sentimentos e, posteriormente, através de um modelo de classificação que tente prever os comentários com propensão de serem úteis. Isso permitirá às plataformas de recomendações, como a Yelp, saberem de antemão quais os comentários que serão úteis sem terem de esperar pelos votos dos utilizadores (Bakhshi *et al.*, 2015), otimizando-se assim o processo de recomendação.

Mais adiante, a dissertação encontra-se organizada da seguinte forma: No capítulo 2 é apresentada a revisão da literatura à volta do *social media*, *text mining* e análise de sentimento e as suas relações com a indústria do turismo. Os detalhes da metodologia, explicando cada uma das fases com base nos componentes do projeto, são introduzidos no capítulo 3, enquanto no capítulo 4 é apresentada a análise dos resultados obtidos. A conclusão é feita no capítulo 5, onde também serão indicadas algumas limitações e dificuldades encontradas, bem como algumas propostas para o trabalho futuro.

2 Revisão da Literatura

2.1 *Social Media*

O mundo tem acompanhado de perto a revolução da *Internet* e o acesso, cada vez mais fácil, ao mundo virtual. Em 2004, a empresa americana O'Reilly Media lançou a *Web* 2.0 marcando uma nova era das tecnologias *web* mais voltada às comunidades e aos serviços (Berthon *et al*., 2012). Ao contrário da *Internet* (*Web* 1.0), onde a informação só podia ser consultada, a tecnologia *Web* 2.0 permite a alteração e partilha de informação por meio de *blogs, wikis, youtube* e mundo virtual, o que tem transformado a forma como a empresa e o consumidor interagem *online*. O seu principal objetivo é tornar o ambiente *online* mais dinâmico e fazer com que os utilizadores colaborem na organização de conteúdos através das suas ações individuais (Power & Phillips-Wren, 2011). Isto fez com que muitos *sites* deixassem de ter uma estrutura rígida e estática e tornarem-se em plataformas dinâmicas onde as pessoas podem contribuir com o seu conhecimento para o benefício de outros utilizadores e visitantes.

O surgimento da tecnologia *Web* 2.0 permitiu o aumento significativo da velocidade e facilidade de utilização de muitas aplicações, o que contribuiu para o aumento de conteúdos existentes na *Internet*. Esta tornou-se na maior forma de comunicação existente nos dias de hoje, onde as pessoas criam e partilham conteúdos nas suas redes de comunidades virtuais a um ritmo prodigioso (Asur & Huberman, 2013; Kaplan & Haenlein, 2010), o que tem provocado mudanças abrangentes e significativas na forma de comunicação entre organizações, comunidades e pessoas (Kietzmann *et al*., 2011). A *Web* 2.0 apresenta uma infraestrutura técnica que permite o fenómeno social dos meios de comunicação coletivos e facilita conteúdo gerado pelo consumidor no qual o *social media* é visto como um repositório de conteúdos e o consumidor, o criador e gestor desses conteúdos. Esta disponibiliza um ambiente de interação e participação revelando o termo "s*ocial media"*, definida como um conjunto de aplicações com base na *Internet* construída sobre as bases ideológicas e tecnológicas da *Web* 2.0 que permite de forma descentralizada, a criação e partilha de conteúdos gerados pelo utilizador (Kaplan & Haenlein, 2010), a interação social e a abertura ao público em geral (Abrahams *et al*., 2012). Para os autores Berthon *et.al.* (2012) a definição de *social media* é algo ainda ambíguo e que não tem a aprovação da maioria dos investigadores da área, advogando que a definição indicada acima é muito genérica, gerando divergência no que respeita ao âmbito e o significado do termo.

Na origem desta divergência estão as fontes do *social media*, por estas serem capazes de mudar os conteúdos rapidamente e com pouco controle sobre a sua precisão. Neste sentido, a Wikipedia em 2011 (citado por Cazzava, 2015) apresentou a sua definição defendendo que os *social media* são *medias* voltados para a interação social, usando alta escala de acessibilidade às técnicas de comunicação através da utilização de tecnologias *web* e móvel para transformar a comunicação em diálogo interativo. Por sua vez, Berthon *et al*. (2012) resumem o *social media* ao uso de tecnologias baseadas na *web* para melhorar a comunicação humana e criar diálogos dinâmicos e interativos. Mas

ainda há quem defenda que é praticamente impossível diferenciar o *social media* da *web* e, para muitos, falar de um ou de outro é referir-se ao mesmo (Cavazza, 2015). Apesar de não ser consensual a definição entre os vários autores, o *social media* é sem dúvida consequência da *Web* 2.0.

O *social media* tem mudado com o avanço das tecnologias, tendo cada vez mais capacidade para armazenar o conteúdo gerado pelo utilizador, suportando vários tipos e tamanho de dados e facilitando a sua distribuição. Neste ambiente os utilizadores podem combinar, editar e arquivar conteúdos facilmente. A publicação de ideias e opiniões não é avaliada, censurada ou validada (Berthon *et al*., 2012), os utilizadores são livres para escreverem o que quiserem, quando quiserem como quiserem. Este é um fenómeno que tem preocupado muito as empresas porque os consumidores são cada vez mais livres e criativos (Berthon *et al*., 2012), sendo cada vez maior o número de aderentes aos vários tipos de *social media* existente, onde podem positiva e negativamente afetar o raciocínio e eficácia das tomadas de decisões uns dos outros.

Kaplan e Haenlein (2010) identificaram seis diferentes tipos de *social media*: projetos colaborativos, *blogs* e *microblogs*, comunidade de conteúdos, *sites* de redes sociais, mundo de jogos virtuais e sociedade virtual. Por sua vez, Cavazza (2011) dividiu o *social media* em 10 categorias alterando ligeiramente a estrutura apresentada pelos autores anteriores, onde figura a: publicação (*Wikipedia*), partilha (*SoundCloud*), discussão (*Quora*), rede social (*Google+), microblog* (*Pownce*), *lifestream* (*friendfeed*), *livecast* (*justin.tv*), mundo virtual (*Second Life*), jogos sociais (*Click* Jogos) e jogos *online multiplayer* (*Wizard 101*). Mas a categorização do *social media* vem sofrendo alterações ao longo dos anos tornando-se mais convergente, pois há já muitas áreas diferentes que integram tecnologias que lhes permitem fazer de tudo um pouco dentro de um tipo específico de *social media*. Neste sentido, Cavazza reestrutura, todos os anos, o panorama do *social media* que ele próprio o apresenta. A Figura 1 mostra que, no presente, o *social media* se encontra dividido em apenas quatro categorias distintas com os diversos serviços que cada um disponibiliza.

Figura 1 - O Panorama do *social media* em 2015

Fonte: Cavazza, 2015

Os diferentes canais de *social media* continuam a crescer e a cada dia que passa surgem novos. A Figura 1 apresenta apenas alguns deles, com destaque para as plataformas dominantes: o Facebook e o Twitter no centro da estrutura. Não porque são melhores do que os outros, mas sim porque eles são a extremidade da cadeia onde se concentram e se transmitem todas as interações sociais que são feitas nas outras plataformas. Estes não seriam tão poderosos sem o conteúdo publicado e partilhado por todo o *social media*. Na segunda camada encontram-se algumas aplicações móveis como o Viber, o Messenger e o WhatsApp. Estas aplicações suportam várias funcionalidades como comprar produtos, chamar um táxi, trocar dinheiro e muitas outras funcionalidades que já não são suportadas pelo Facebook ou Twitter. Na camada seguinte encontram-se vários serviços divididos em quatro grandes áreas: a publicação, a partilha, a discussão e a rede. Estas áreas são complementares, pois o utilizador publica um conteúdo, partilha com os outros e gera conversas que lhe permitem desenvolver a sua própria rede de contactos. Foi esta interação que fez criar o interesse das empresas nos *social media*, principalmente nas empresas ligadas ao turismo onde o conteúdo gerado pelo utilizador é a chave de garantia e qualidade dos seus produtos (Litvin *et al.*, 2008).

Um canal muito utilizado neste processo de interação são os *sites* de recomendação onde, por exemplo, um turista, depois de passar as suas férias num determinado lugar, publica a sua experiência de contacto e consumo dos vários produtos oferecidos durante as férias, deixando fotos, comentários, classificação e uma série de informações que podem ser úteis para os próximos turistas, que são os

clientes e potenciais clientes da indústria turística. Os *sites* de recomendação são um tipo de *social media* pertencente à área de redes, pois os consumidores acabam por construir as suas próprias redes de interesse na plataforma.

2.1.1 A importância do *social media* para o marketing

Os consumidores, que estão cada vez mais atentos, são adeptos ativos das novas ofertas tecnológicas, procurando por melhores oportunidades de compra, baseando-se nos comentários feitos por outros consumidores que partilham as suas experiências reais de relacionamento com os produtos, serviços, empresas, etc. Esta partilha é de grande utilidade para outros consumidores, ajudando-os a decidir onde gastar o seu dinheiro (Zeng & Gerritsen, 2014).

Uma das principais características do *social media* é a transmissão de informação de forma rápida e abrangente, alcançando um grande número de pessoas em frações de segundo e tornando-se numa das mais poderosas ferramentas de consultas, solicitações, reclamações, recomendações, entre muitas outras formas de expressões, pois muitos são os consumidores que se baseiam nas informações apresentadas num determinado *social media* para tomarem as suas decisões.

Por isso, o *social media* tornou-se num dos principais influenciadores do processo de decisão de compra dos consumidores por permitir a construção de diversas comunidades *online* que constituem grupos de referência para os consumidores. O principal papel destes grupos é permitir a partilha de experiência que permite esclarecer as dúvidas do consumidor e ajudá-lo a decidir. São vários os fatores internos e externos que influenciam o comportamento do consumidor no processo de tomada de decisão de compra. Entre elas encontra-se os fatores culturais (cultura e classe social), os fatores sociais (grupos de referência, família, regras e estado), os fatores pessoais (idade, circunstância económica e ocupação), os fatores psicológicos (motivação, aprendizagem, perceção e atitude) e o fator comprador (quem irá efetuar a compra) (Gillingan & Wilson, 2003, pp. 225-234).

O marketing, tendo o seu propósito centrado na satisfação das necessidades e desejos dos consumidores tem de estar atento a estes influenciadores e comparar com a informação que tem sobre o comportamento de compra dos consumidores para ganhar conhecimento e poder agir. Com foco em diferentes papéis de compra, no tipo de comportamento e no processo de decisão, a estratégia de marketing estará em posição de examinar o processo de compra em si. Estes três elementos são melhor explicados por Gillingan e Wilson (2003, pp. 235-243) onde identificam cinco papéis distintos no processo de compra:

1. **O iniciador** – quem inicialmente sugere a compra do produto ou serviço;
2. **O influenciador** – cujos comentários afetam a tomada de decisão;
3. **O decisor** – quem por último toma toda ou parte da decisão de compra;
4. **O comprador** – quem fisicamente compra;
5. **O utilizador** – quem consome o produto ou serviço.

Quanto aos diferentes tipos de comportamento de compra, apresentam a compra habitual, a procura de variedade, a redução de dissonância e compra complexa como base no custo,

complexidade, risco e oportunidade da decisão de compra. E, por último, para perceber "como o consumidor compra um produto em particular", os autores apresentaram uma série de modelos que não centram apenas sobre a decisão de compra, mas também sobre o processo que conduz a essa decisão, a decisão em si, e posteriormente o comportamento pós-compra. Um exemplo do tipo de modelo é apresentado na Figura 2.

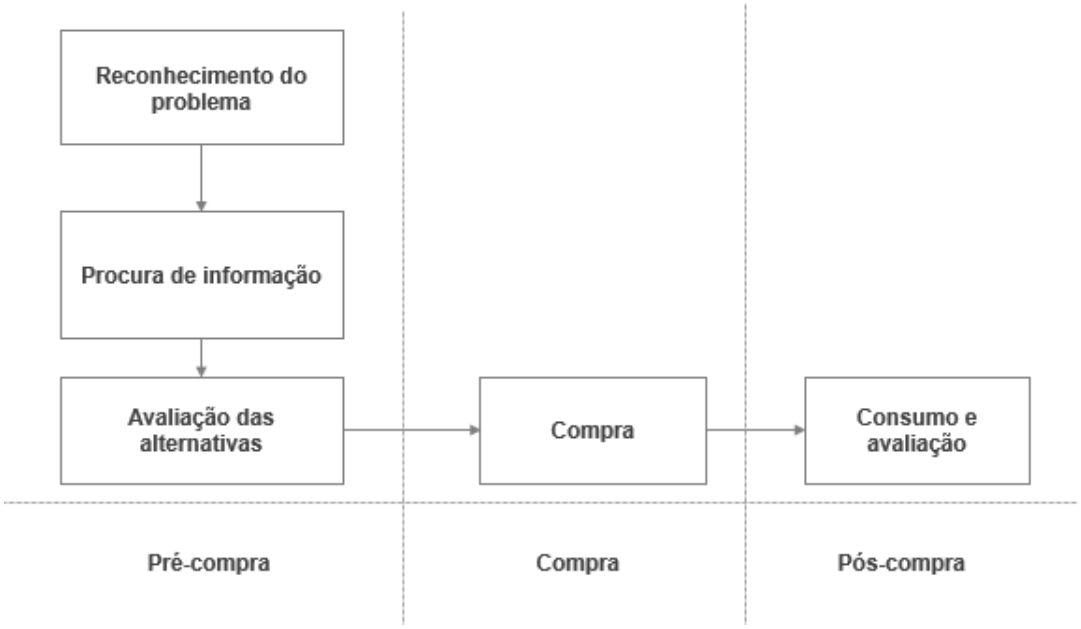

Figura 2 - Modelo sequencial do processo de compra

Os modelos apresentados foram:

- **O modelo de Nicosia** – Procura explicar como um potencial consumidor responde à notícia de uma nova marca. Centra-se no fluxo de informação entre a empresa e os consumidores, e na forma como a empresa exerce a sua influência sobre os consumidores e como estes influenciam aquela empresa. Para isso, tem em conta o efeito de três fatores: a atitude, a motivação e a experiência.
- **O modelo de Howard e Seth** – Estuda o comportamento de compra do consumidor partindo da eleição da marca.
- **O modelo de Engel, Kollat e Blackwell** – Descreve, de forma geral, o processo de compra e as relações entre as variáveis que nele intervêm. Tenta explicar a perceção que o consumidor tem da necessidade que deve ser satisfeita.

Os modelos de comportamento do consumidor proporcionam uma imagem global sobre os comportamentos dos consumidores nos vários momentos de compra e variáveis que a influenciam e permitem identificar as áreas e variáveis-chave que devem ser tidas em conta na tomada de decisão do marketing. Isto é útil porque o consumidor já não é um ser indefeso, mas sim informado e formado, com capacidade de exigência perante o seu meio e face às diferentes influências que encontra.

De acordo com a Figura 2, quando um consumidor vai à procura de informação nos *social media,* como as plataformas de recomendações para obter informação que o ajude na sua tomada de decisão, ele se encontrará na etapa de "Pré-compra" na fase "Procura de informação" que depois, mediante a informação encontrada, vai poder avaliar as diversas alternativas. Aqui o *site* de recomendação, como já foi dito anteriormente, é o influenciador do consumidor no processo de compra. As influências que o consumidor encontra através dos diferentes tipos de *social media* são provenientes de um grande volume de dados de conteúdo e qualidade variável gerados por outros consumidores. Navegar nesse conteúdo é um desafio quando se pretende levar a cabo uma investigação significativa (Parameswaran & Whinston, 2007; Gopal *et al*., 2011), que pode exigir aplicação de filtros, conteúdo semântico, *tagging*, *information mining* ou outras técnicas (Abrahams *et al*., 2012). Para as organizações, o desafio centra-se no acesso à informação escondida nestes conteúdos e a sua transformação em algo que lhes seja útil. Tratando estes dados será possível entender as necessidades e comportamentos dos clientes ou grupos específicos, relativamente aos produtos e serviços de uma dada organização. Perceber o que as pessoas gostam ou não, e o que pensam sobre a organização em si e os seus concorrentes, é a chave para o sucesso de negócio de qualquer organização.

Por isso, há um número cada vez maior de empresas que estão fazendo uso destas ferramentas (Leung *et al*., 2013) para anunciarem os seus produtos e divulgarem informações aos seus *stakeholders* (Asur & Huberman, 2013). Além disso, a informação gerada é utilizada para conhecer melhor os clientes e mais consumidores, o que ajudará a empresa a desenvolver planos de marketing individuais ou personalizados para os atuais e potenciais clientes (Miller, 2005, pp. 26).

Conscientes de que "o cliente é a razão de ser de qualquer organização" (Lindon *et al*., 2009, pp.633), as empresas estão cada vez mais empenhadas em desenvolver estratégias que incorporam o relacionamento com os seus clientes. Um dos pontos desta estratégia é a implementação do *Customer Relationship Management* (CRM) que, segundo Greenberg (2010), é:

> *"uma estratégia de negócios para selecionar e gerir o cliente no sentido de otimizar o seu valor a longo prazo. O CRM requer uma filosofia e cultura empresarial centrada no cliente para dar suporte de forma eficaz aos processos de marketing, vendas e serviços. As aplicações de CRM podem ajudar na gestão da relação com o cliente tornando-a mais eficaz, desde que a organização possua a liderança, as estratégias e a cultura certa."*

Com a implementação desta estratégia, a empresa pode realizar iniciativas de marketing para os seus atuais e futuros clientes que visam atingir os consumidores certos a fim de maximizar o *Consumer Lifetime Value*. Esta abordagem visa adquirir novos clientes, mas sobretudo fidelizar os atuais e torná-los mais valiosos, através de compras repetidas, usando-se o conhecimento dos clientes para aumentar a sua satisfação, incrementar a venda cruzada de produtos (*cross-selling*) e aumentar a venda de produtos mais caros (*up-selling)* (Lindon *et al*., 2009, pp.637). Independentemente do produto, do serviço ou da marca, existe a clara noção de que quanto mais forte for a relação

estabelecida com o consumidor, mais forte é o vínculo à marca no curto ou longo prazo, com tradução direta na rentabilidade do cliente.

Os *social media* contribuem bastante para o desenvolvimento do CRM, uma vez que permitem recolher informações contínuas sobre os clientes, ter elevada interatividade com eles, estabelecer comunicação através das diferentes redes sociais existentes (inclusive *sites* de comentários *online* como a yelp.com), estar perto do cliente em qualquer momento e em qualquer lugar, e comunicar de forma individualizada com cada um deles; em suma, o *social media*, tornou-se num dos principais meios de comunicação utilizada pelos profissionais do marketing para chegar aos consumidores em geral. Estas são as principais janelas de oportunidade aproveitadas pelo *social media* comparativamente a outras estratégias de marketing.

O CRM tradicional tornou-se incapaz de cumprir com todos os requisitos necessários para uma gestão eficaz da relação da organização e dos seus clientes, uma vez que é apenas um sistema de repositório de dados transacionais que regista, por exemplo, o que o cliente compra, a que oportunidades de *up-selling* e *cross-selling* o cliente pode responder, ou como ficou resolvida uma determinada reclamação do cliente. Este sistema fornece um nível de perceção limitado e focado apenas no conhecimento comportamental, com indicadores de satisfação e/ou lealdade também limitados, sendo que a maioria dos dados presentes no sistema de CRM são recolhidos independentemente da participação direta do cliente (Greenberg, 2010). É esta participação do cliente que faz a diferença, pois estes estão cada vez mais exigentes e inteligentes. Para poder dar resposta às suas exigências foi desenvolvido o CRM 2.0, também conhecido por *social* CRM (Greenberg, 2010), que permite extrair informações *online* geradas pelos clientes e com elas elaborar estratégias de marketing que vão ao encontro das reais necessidades dos mesmos.

Greenberg (2010) defende que o CRM 2.0 é a resposta da empresa para fazer face à nova geração de clientes (participativos, criativos e dependentes do *social media*). Define-o como uma filosofia e uma estratégia de negócio, suportadas por uma plataforma de tecnologia, regras de negócio, processos e características sociais, projetados para envolver o cliente numa comunicação colaborativa, a fim de fornecer o valor mutuamente benéfico num ambiente de negócio confiável e transparente. Para que seja possível ter uma melhor perceção sobre o cliente, Greenberg (2009) aconselha a apropriação de cinco componentes essenciais para adquirir o tipo de informação necessária para melhorar a capacidade de aprendizagem de todo o negócio em torno do cliente:

- **Dados:** Além dos recolhidos do sistema CRM, tem-se dados do registo de cliente e dados externos recolhidos dos blogs, redes sociais, *sites* de recomendação, etc, que são analisados através do *text mining* . Isso acrescenta uma dimensão importante para os dados transacionais mais estáticos e informações corporativas.
- **Análise de sentimentos:** Escrutínio usado para medir a temperatura emocional de indivíduos e grupos. Ele analisa as atitudes positivas, negativas ou neutras dos clientes num momento em particular ou mudanças de atitude ao longo do tempo. Esta análise também pode ser utilizada para medir a forma como os efeitos de uma certa atitude propaga através das redes sociais ou comunidades. O sentimento da mensagem é

identificado, medido e classificado (Stieglitz & Dang-Xuan, 2012). Extrair sentimento dos textos é uma funcionalidade de grande importância para qualquer negócio que lida constantemente com um grande volume deste tipo de dados.

- **Monitorização do *social media*:** O *social media* fornece uma rica fonte de informação. O interesse que os utilizadores individuais revelam nas suas opiniões *online* sobre produtos e serviços e a potencial influência que tais pareceres exercem é algo a que as empresas estão a dar mais e mais atenção (Pang & Lee, 2008). Os profissionais de marketing precisam de monitorizar sempre os *social media* a fim de obterem informações atualizadas relacionadas com as suas marcas. Mas esta monitorização não é possível de ser realizada através de métodos tradicionais, devido ao fato de o *social media* ser fragmentado e o comportamento do utilizador estar sempre em mutação (Pang & Lee, 2008). Neste sentido, tem crescido o interesse em adotar sistemas que sejam capazes de analisar automaticamente interesses do consumidor que são expressos, em grande parte, no mundo virtual, possibilitando às empresas entender como os seus produtos e serviços são percecionados (Pang & Lee, 2008).
- **Perfis:** Esta é a informação que caracteriza o indivíduo. Esta informação é importante para obter uma visão do cliente e da forma como este quer interagir com a empresa. Isso pode incluir interesses de cinema, literatura, *hobbies* e gostos, tudo extraído do texto não estruturado que ele publicou. Com o crescente interesse em micro-segmentação, o mergulho profundo na vida do cliente, sem ser invasivo, permite entender as suas escolhas de estilo e de seleção a fim de prever futuros comportamentos aparentemente não relacionados. Os perfis tornam-se essenciais para encontrar informações diferenciáveis sobre o cliente.
- **Mapas de experiências do cliente:** O mapeamento da experiência do cliente promove o conhecimento sobre o mesmo, uma vez que transpõe o conhecimento geralmente incorreto que se tem sobre o que o cliente pensa (no sentido de dar a dimensão emocional que permite conhecer as atitudes do cliente?).

Dos componentes apresentados acima o presente projeto aborda os dois primeiros, Dados e Análise de sentimentos. Sobre os dados extraídos do *social media* é possível aplicar métodos preditivos que permitem antecipar comportamentos e, com isso, desenvolver estratégias de Marketing Relacional com efeito preventivo, para evitar comportamentos que prejudicam a empresa (como a perda do cliente para a concorrência). O marketing relacional e o CRM são nomes diferentes, mas que traduzem o mesmo conceito, apesar de muitas vezes este último ser confundido com os programas que lhe dão suporte (Lindon *et.al.,* pp.641). O marketing relacional é uma forma de atrair, manter e aumentar (em organizações de multi-serviços) as relações com os clientes, onde mais do que atrair novos clientes o importante é manter os clientes atuais, fidelizando-os de forma mais lucrativa (Berry, 2002). O CRM faz parte desta estratégia de marketing que implica ter uma visão única e integrada do cliente, compreendendo “todas as operações, processos e tecnologias que se desenvolvem tendo como

objetivo o cliente e a satisfação das suas necessidades, numa atitude pró-ativa da empresa" (Ferrão, 2003).

A w*eb* desempenha um papel muito importante no marketing, uma vez que esta mudou a relação existente entre Publicidade, Gestão da Marca e Marketing Direto, possibilitando saber com algum grau de precisão quem está a ver e a aceder a cada anúncio, e o que leva o consumidor a comprar (Linoff e Berry, 2001). A análise dos dados provenientes da relação com os clientes é realizada no âmbito do CRM analítico, que permite analisar os dados dos diferentes pontos de contacto da empresa com o cliente. Neste projeto, o ponto de contacto utilizado para a análise de dados é o *site* de recomendação "http://www.yelp.com", de onde são recolhidos comentários feitos pelos diferentes consumidores sobre os vários ramos de negócio dentro da indústria turística.

A *Yelp* detém um *site* de comentários de negócios organizado em torno de proximidades geográficas, onde se encontram comentários relativamente às empresas locais, como restaurantes, livrarias, museus, etc., publicados por consumidores que tiveram uma certa experiência com estas empresas. Estes caracterizam de certa forma o produto ou serviço em questão e podem afetar outros consumidores nas suas decisões de compra (Greenberg, 2010; Guernsey, 2000; Zeng & Gerritsen, 2014). Anteriormente, essas decisões baseavam-se em anúncios ou informações sobre produtos fornecidos pelos vendedores. No entanto, com a proliferação de *e-commerce* e o aumento do número de comentários sobre produtos fornecidos pelos consumidores, verificou-se que os consumidores têm invocado comentários *online* na busca de informações relacionadas com uma variedade de produtos (Hu *et.al.,* 2012).

Uma vez que os comentários *online* têm vindo a conquistar terreno e a ganhar credito no meio das comunidades virtuais (Hu *et.al.,* 2012), já não é suficiente para os negócios confiar apenas nos meios tradicionais para o desenvolvimento do marketing. Tal fato tem atraído a atenção destes profissionais no sentido de desenvolverem marketing no mundo virtual (Dellarocas, 2003). Isto faz do *social media* uma estratégia de marketing ao atuar como um canal de divulgação deste tipo de informação. Um fator muito importante a ter em consideração no marketing *online* é o conteúdo gerado pelo utilizador, também conhecido como o WOM virtual (Litvin *et.al.,* 2008).

Entre todas as fontes de informação disponíveis, o WOM tem sido reconhecido por muito tempo como uma das importantes fontes de informação externas devido à sua alta credibilidade junto dos consumidores (Murphy *et.al.,* 2007). Este tornou-se num veículo muito poderoso, utilizado para transportar as mensagens do marketing (Kozinets *et.al.,* 2008). Os consumidores muitas vezes reveem-se nos comentários dos outros e compreendem os produtos com base na perceção de seus "amigos" ou colegas consumidores (Leung *et.al.,* 2013). Esta forma de transmitir informação tem ganho adeptos que tendem a ser mais influenciados desta forma do que pelas fontes comerciais. Por isso é que o *social media* se tem tornado uma ferramenta muito importante para a estratégia quer da organização em si, como do marketing em particular. Estudos mostram que as empresas não só publicam regularmente as suas informações de produtos em fóruns *online* como também incentivam pró-ativamente os seus consumidores a espalhar a palavra sobre seus produtos (Godes & Mayzlin, 2004). Assim, a *social media* pode ser usada como um meio económico, permitindo aos profissionais de

marketing chegar a milhões de utilizadores com apenas uma quantidade insignificante de recursos, reduzindo os gastos com a publicidade (Hochreiter & Waldhauser, 2014).

Utilizar *social media* para fins de marketing é uma escolha defensável, pois trata-se de acompanhar a evolução tecnológica rumo ao crescimento e ser capaz de interagir com os consumidores nos canais normalmente associados a amigos e pessoas interessantes que os seguem. Com isso, constrói-se um contexto de marketing atrativo e isso transforma a presença de marketing em ativação do consumidor (Hochreiter & Waldhauser, 2014). O *feedback* do consumidor fornece uma valiosa fonte de informação para melhorar projetos de produtos e estratégias de marketing (Leung *et.al.,* 2013).

O cliente do século XX tornou-se no cliente social do século XXI. Eles são clientes que se sentem obrigados a partilhar informações com pessoas que provavelmente nunca virão a conhecer, mas que são parecidas com eles. O conteúdo gerado pelo utilizador no *social media* e a sua partilha podem ser vistos por milhões de utilizadores (Greenberg, 2010). Nestes termos, os comerciantes devem ter por objetivo maximizar a partilha das suas mensagens. É útil pensar no utilizador individual como um filtro através do qual uma mensagem tem de passar para chegar a mais utilizadores dentro da rede de utilizadores (Greenberg, 2010).

Esta tendência tem ganho força na indústria turística, onde o *social media* tem sido utilizada para promover as localidades turísticas e as economias locais. Alguns estudos recentes têm revelado que os *social media* vêm desempenhando um papel importante não só para os consumidores em busca de informações, mas também como uma ferramenta de marketing de turismo (Chan & Guillet, 2011; Huang, 2012; Inversini *et.al.,* 2009; Munar, 2011; Xiang & Gretzel, 2010).

2.1.2 *Sites* de recomendação

Esta publicação e partilha de conteúdos são funcionalidades existentes em vários canais e cada vez mais as diferentes áreas têm vindo a implementar esta funcionalidade, permitindo que os utilizadores partilhem seja o que for nos diferentes tipos de *social media* mesmo que sejam opiniões e comentários relacionados com empresas, negócios, produtos ou serviços. Mas quando se pretende encontrar uma dada informação sobre um produto ou uma marca específica o ideal é procurá-los nos canais certos para não se perder no meio da infinidade de informação armazenada nos diferentes *sites* onde, muitas vezes, se acaba por não encontrar o que realmente se precisa.

Então, para reunir as empresas, os diferentes tipos de negócios e os consumidores num só espaço, surgiram os *sites* de comentários *online*, mais especificamente os *sites* de recomendação, tendo como principal objetivo reunir num só espaço as empresas e os consumidores. Por um lado, a empresa pode mais facilmente filtrar os conteúdos que lhe dizem respeito, a fim de analisar a apreciação dos consumidores em relação aos seus bens e serviços para, de seguida, oferecer aqueles que realmente vão de encontro às necessidades do consumidor, por outro lado, o consumidor quando precisar de dicas, recomendações ou fazer comparações, consulta estes tipos de *sites* que se têm revelado num valioso meio de ajuda para os consumidores da *Web,* uma vez que lhes oferecem

recomendações e sugestões úteis e eficazes com base nas experiências de outros consumidores (Tarannum *et.al.,* 2015; Litvin *et.al.,* 2008).

Por este motivo, as empresas ligadas ao turismo estão mais atentas a este canal devido ao facto de os seus produtos carecerem bastante dos fatores garantia de qualidade e segurança, pois são constituídos por propriedades intangíveis (Litvin *et.al.,* 2008). Daí a necessidade de haver quem fale bem dos seus produtos, sendo os turistas os seus melhores aliados. Um exemplo deste tipo de *social media* é o Yelp.com, considerado um dos melhores *sites* para procurar recomendações sobre os melhores negócios de um determinado lugar. Adotando este meio de comunicação com o consumidor, a empresa acumula benefícios, como baixo custo com a publicidade, e transmissão de qualidade, garantia e confiança por parte de consumidores experientes, o que pode resultar na fidelização e angariação de novos clientes (Leung *et.al.,* 2013).

Os consumidores muitas vezes descrevem as suas experiências pessoais e memórias que tiveram na utilização de um dado produto ou serviço, dando às empresas um conteúdo excelente para explicar aos seus próximos e potenciais clientes as razões pelos quais devem adquirir o produto ou o serviço fornecendo uma base importante para a tomada de decisão. Porém, nem sempre esses comentários são de todo favoráveis à empresa, o que pode comprometer a sua reputação e sobrevivência no mercado competitivo (Tarannum *et.al.,* 2015). Por isso, a empresa tem de monitorizar de perto o que acontece no mundo virtual (principalmente nos *sites* de recomendação onde se concentram vários concorrentes num só espaço).

Analisando os comentários feitos em diferentes *sites* de recomendação, as empresas podem adquirir algumas perceções sobre o que os consumidores gostam e não gostam, bem como os *add-ons* que estes estão dispostos a pagar (Leung *et.al.,* 2013). Depois de ter uma rica compreensão das características dos clientes e seus padrões de comportamento, as empresas podem utilizar essa forma de comunicação, denominada *word-of-mouth* (WOM) virtual para elaborar estratégias, a fim de reforçar a proposta de valor e aumentar o patrocínio do cliente (Leung *et.al.,* 2013). Para isso, é necessário a implementação de técnicas de personalização que permitem destacar os conteúdos considerados mais úteis ao consumidor com base nas suas preferências e gostos. Um sistema de recomendação é implementado como a principal ferramenta de personalização. Isto é necessário para corresponder aos potenciais interesses e expectativas dos consumidores (Adomavicius & Tuzhilin, 2005).

A plataforma Yelp.com reúne diversos negócios em torno do turismo presentes num determinado lugar, tendo incorporado um sistema de recomendação que avalia cada comentário do consumidor de forma automática com base na qualidade, fiabilidade e atividade do consumidor na Yelp. O sistema procura comentários cujo conteúdo, positivo ou negativo, possa ser útil e fiável aos consumidores. Com isto, da totalidade dos comentários publicados apenas 75% são recomendados (Yelp, 2015d). O sistema de recomendação está constantemente em execução e a fazer ajustes ao mesmo tempo que recolhe mais informação sobre comentários e consumidores. Assim, os comentários apresentados para cada empresa vão mudando com o tempo (Yelp, 2015d). A Yelp tenta sempre colocar em destaque os comentários que foram considerados pelos consumidores como sendo úteis e de confiança, sendo a confiança um dos fatores-chave para o sucesso da indústria turística.

2.1.3 *Social media* e a sua influência na Indústria Turística

As empresas sentiram que a *social media* lhes traria muitos benefícios para elas se estivessem na mesma comunidade *online* onde os seus consumidores gastam o seu tempo, podendo partilhar as suas novidades, publicitar os seus produtos e serviços e interagir com os seus clientes, incentivando-os a partilharem as suas experiências por meio de blogs, redes sociais, *sites* de recomendação, entre muitos outros espaços *online*.

O *social media* vem desempenhando um papel importantíssimo na indústria do turismo especialmente no que diz respeito à promoção do turismo, à procura de informações e comportamentos do consumidor relativamente à tomada de decisões (Fotis, 2012), foco nas melhores práticas de interação e relacionamento das empresas turísticas com os consumidores (Zeng & Gerritsen, 2014), e tem sido uma excelente estratégia na promoção dos produtos turísticos (Fotis, 2012).

O turismo é visto como um fenómeno social, cultural e económico que implica a deslocação e permanência de pessoas para países ou lugares fora do seu ambiente habitual, não mais de que um ano consecutivo, para fins pessoais, de negócios, profissionais ou outros (UNWTO, 2015a). Essas pessoas são chamadas visitantes, mas geralmente são tratadas por turistas. O Turismo resume-se num conglomerado de atividades dos visitantes (IRTS, 2008, 2.9).

A *United Nations World Tourism Organizations* (UNWTO) é uma organização que promove o turismo como um motor de crescimento económico, desenvolvimento inclusivo e sustentabilidade ambiental, e oferece liderança e apoio à indústria a nível de conhecimentos avançados e políticas de turismo (UNWTO, 2015b). Segundo esta organização, o turismo é uma indústria que tem vindo a crescer e a diversificar-se ao longo dos tempos no sector económico em todo o mundo. Este encontra-se intimamente ligado ao desenvolvimento e o número de destinos turísticos tem aumentado consideravelmente, sendo a *social media* um dos responsáveis por este crescimento (Zeng & Gerritsen, 2014). Esta dinâmica alavancou o turismo tornando-o um motor essencial no progresso socioeconómico (UNWTO, 2015c).

No presente, o volume de negócios do turismo iguala ou até supera o volume das consideradas principais potências económicas mundiais, como a exportação do petróleo, produtos alimentares e automóveis. A construção do turismo para o bem-estar económico depende da qualidade e das receitas da oferta turística, mas não é fácil para os profissionais da indústria turística transmitir e garantir a qualidade dos seus produtos por serem algo intangível (Litvin *et.al.,* 2008). Uma forma de diminuir esta dificuldade é a aproximação das empresas do turismo ao consumidor através do *social media*, uma estratégia que tem tido ótimos resultados, pois através deste meio as empresas chegam mais facilmente aos seus consumidores e contam com eles para partilharem as suas experiências que são, muitas vezes, informações úteis e seguras quanto à qualidade dos produtos oferecidos.

Estando perante uma situação de "Comprar" ou "Não comprar", o consumidor, muitas vezes fica apreensivo, pois sabe que está a correr um risco, uma vez que nem tudo o que vê nas publicidades corresponde à realidade. E como diz o ditado, muitas vezes não se sabe se não se está "a comprar gato por lebre". Por esta razão, as pessoas querem sempre saber mais informação sobre o que tencionam comprar e quanto maior for a sensação de risco no contexto pré-compra, maior a propensão

para a "caça" de informações seguras e confiáveis. Um turista, por exemplo, procura por informações sobre destinos de viagens e recursos a elas associadas, tais como alojamentos, restaurantes, museus, eventos, entre muitos outros serviços, a fim de planear uma viagem (Borràs, *et.al.,* 2014).

Nesta indústria, a necessidade de informação é enfatizada por certas características do produto turístico, entre elas a intangibilidade, onde o produto não pode ser inspecionado antes da compra. É impossível fornecer uma amostra do produto ao turista, que não tem como comparar os produtos que irá usar, a não ser no momento do consumo. Por isso é que muitos autores defendem a necessidade de os turistas se sentirem assegurados, procurando produtos que já venham recomendados e bem referenciados por outras pessoas com experiência de utilização (Litvin *et.al.,* 2008; Leung *et.al.,*2013; Fotis, 2012; Munar & Jacobsen, 2014).

Através dos diversos tipos de *social media*, os turistas partilham as suas experiências de viagens, e esta partilha é reconhecida como uma importante fonte de informação que contribuirá na planificação das viagens turísticas ou influenciará potenciais viajantes a tomarem as suas decisões (Zeng & Gerritsen, 2014). A maioria dos profissionais de turismo que anunciam o seu negócio *online,* permitem e encorajam os turistas a deixar comentários e recomendações sobre os seus produtos, serviços e experiências. Os turistas gostam e necessitam de ser assegurados da qualidade do serviço que estão a adquirir. Daí, com frequência, um testemunho de um turista prévio pode ser bem mais poderoso do que as publicidades e anúncios publicados (Litvin *et.al.,*2008). Sendo assim, o turista é quase totalmente dependente de representações e descrições para o ajudar na sua decisão. Dessa forma, o acesso a informações precisas e confiáveis é vital para orientá-lo numa escolha adequada.

O Consumer Reports[1], a Yelp[2], o TripAdvisor[3] e o Trivago[4] são alguns dos *sites* de comentários *online* existentes que fornecem dicas e recomendações de vários produtos e serviços com o objetivo de apoiar o consumidor a fazer a melhor escolha possível, com base nos comentários de outros consumidores que já experimentaram determinados produtos ou serviços. Estes comentários são vistos como o WOM virtual (Litvin *et.al.,* 2008), representando uma ótima oportunidade para as empresas promoverem os seus respetivos produtos, criando para si vantagens competitivas face aos concorrentes (Litvin *et.al.,* 2008). É uma verdadeira fonte de informação para o consumidor que pode comparar produtos e/ou serviços, ao nível de qualidade, preço e concorrência, fazer uma avaliação antes de decidir o que consumir e, finalmente, decidir de que empresa se tornará cliente.

A Consumer Reports é uma organização independente sem fins lucrativos que visa apoiar o consumidor através de testes e avaliações imparciais do produto, pesquisa, jornalismo, educação pública e advocacia. Permite que o consumidor possa ter uma verdadeira visão do mercado e fazer as suas comparações antes de tomar uma decisão. A Consumer Reports funciona como uma entidade de apoio ao consumidor defendendo o princípio de que os produtos e serviços devem ser seguros, eficazes, de confiança e a preços justos, insistindo com os fabricantes, retalhistas, agências governamentais e outros para que sejam claros e honestos. Realizam inquéritos anuais a mais de um

[1] http://www.consumerreports.org/, acedido em 30-03-2015
[2] http://www.yelp.com, acedido em 13-02-2015.
[3] http://www.tripadvisor.com, acedido em 13-02-2015.
[4] http://www.trivago.com, acedido em 13-02-2015.

milhão de assinantes a fim de captar o *feedback* vital nas decisões, experiências e hábitos de compras que os ajuda a aprofundar o trabalho de informação e proteção. O seu objetivo passa pela orientação, informação e sensibilização dos consumidores, para que estes possam tomar decisões estando muito bem informados (Consumer Reports, 2015).

Por sua vez, o TripAdvisor é o maior *site* de viagens do mundo, permitindo que os viajantes planeiem e reservem a viagem perfeita. O TripAdvisor oferece dicas de milhões de viajantes e uma grande variedade de opções de viagem e recursos de planeamento, com *links* integrados com ferramentas de reservas que verificam centenas de *sites* para encontrar os melhores preços de hotéis. Os *sites* com a marca TripAdvisor formam a maior comunidade de viagens do mundo, chegando a 375 milhões de visitantes únicos mensais e gerando mais de 250 milhões de comentários e opiniões que cobrem acima de 5,2 milhões de acomodações, restaurantes e atrações. Os *sites* operam em 45 países em todo o mundo. O TripAdvisor também inclui TripAdvisor for Business, uma divisão se dedica a oferecer à indústria turística o acesso aos milhões de visitantes mensais do TripAdvisor. A *TripAdvisor*, administra e opera *sites* sob outras 24 marcas de *media* relacionadas com viagens (TripAdvisor, 2015).

Restrito a um único sector de atividades, o Trivago é considerado o maior motor de busca de hotéis do mundo e compara diariamente 858.247 hotéis de 266 *sites* de reserva. A sua missão é ser e manter-se como a primeira fonte de informação independente dos viajantes para encontrar o hotel ideal ao melhor preço. Detém cerca de 52 plataformas em mais de 30 idiomas onde se pode verificar os resultados da fusão de tecnologia de ponta com um *design* apelativo e um conteúdo de hotéis extraordinário (Trivago, 2015).

Por seu turno, a Yelp procura indicar e recomendar os comentários que sejam de grande utilidade para o consumidor, ajudando-o a encontrar os melhores negócios locais, como um bom restaurante ou um dos melhores hotéis (Yelp, 2015a). Estes comentários são fornecidos pelos vários consumidores que tiverem alguma experiência com uma determinada empresa, tendo em conta os seus produtos e serviços. Em 2014, este *site* registou cerca de 71 milhões de comentários e, através da utilização de um *software* automático, a Yelp recomenda os que são considerados úteis e confiáveis para a sua comunidade, selecionando-os de entre os milhões que recebe. O *software* analisa dezenas de sinais diferentes, incluindo várias medidas de qualidade, confiabilidade e atividade do comentador no *site*. E para que as empresas locais saibam como lidar com estes comentários, a Yelp promove formação que ensina a responder a estes comentários de forma responsável (Yelp, 2015d).

De acordo com a Alexa (2015d), tendo em conta a popularidade relativamente a outros sites de comentários *online* acima apresentados, a Yelp está classificada na 34ª posição nos Estados Unidos e 150ª posição a nível global desde 23 de julho de 2015. Esta apresenta melhor classificação em relação aos seus concorrentes: a Consumer Reports está classificada na 387ª posição nos Estados Unidos e na 1632ª a nível global (Alexa, 2015a); O Trivago está classificado na 863ª posição nos Estados Unidos e na 3589ª a nível global (Alexa, 2015c) e o TripAdvisor está classificado na 51ª posição nos Estados Unidos e na 175ª a nível global (Alexa, 2015b). O presente projeto recai sobre o *site* yelp.com, por este não ser específico de um produto turístico em particular, mas sim porque é um espaço onde se concentram comentários à volta de todos os produtos relacionados com o turismo.

2.2 *Text Mining*

Considerada uma parte integrante de um conceito mais amplo - *Text Analytics* - o *Text Mining* foca principalmente a descoberta de conhecimentos novos e úteis de informação não estruturada (Sharda *et.al.,* 2014, pp.229). Este é definido como um processo semiautomático de extração de padrões interessantes e não triviais de grandes quantidades de dados textuais não estruturados, de modo a se conseguir um formato estruturado (Miller, 2004). O *text mining* desenvolve um processo muito parecido com o *data mining,* abrangendo os mesmos propósitos, e diferenciando-se apenas no tipo de dados a processar. Em *data mining* são analisados dados estruturados e armazenados em base de dados, enquanto em *text mining* tem-se uma coleção de dados não estruturados (Tan, 1999; Sharda *et.al.,* 2014, pp.230).

A maioria dos trabalhos desenvolvidos no âmbito da descoberta de conhecimento concentram-se em bases de dados estruturadas (Feldman et al., 1998), ignorando os dados não estruturados, pois estes são difíceis de representar e de ser processados automaticamente (Mostafa, 2013). No entanto, estes últimos expressam uma quantidade vasta e rica de informação de grande utilidade para as empresas, apesar do seu formato ser difícil de decifrar automaticamente. Os textos possuem uma estrutura linguística destinada ao consumo dos humanos e não do computador (Provost & Fawcett, 2013, p.252), pelo que esta será a razão pela qual o desenvolvimento dos trabalhos na área de *text mining* apenas ter aparecido muito mais tarde, face ao do *data mining* (Hearst, 1999).

O *text mining* vem ganhando espaço nos últimos anos como uma importante área de descoberta de conhecimentos em dados não estruturados, onde o seu objetivo passa pela organização de enormes quantidades deste tipos de dados que se encontram disponíveis dentro e fora das organizações, como os provenientes dos *sites* da *web*. O *text mining* utiliza-os para obter conhecimento capaz de resolver problemas do mundo real (Godbole *et.al.,* 2010) e oferece inúmeros benefícios às organizações, principalmente para as que lidam diariamente com grandes quantidades de dados textuais. Um exemplo muito interessante é partilhado por Sharda *et.al.,* (2014, p. 230) no qual se usa um formulário de escrita livre dado a um cliente, como os que se encontram em muitos serviços de atendimento ao cliente, dizendo, por exemplo: “queremos melhorar, deixe-nos a sua opinião”. O cliente expõe por escrito e pelas suas próprias palavras, o que pensa sobre os produtos e serviços da empresa em causa ou até mesmo a sua opinião sobre a empresa em si. Com o *text mining* é possível analisar o conteúdo deste formulário ou de qualquer outro documento textual (Abrahams *et.al.,* 2012).

O *text mining* também possibilita a descoberta de padrões em *sites* a fim de extrair dados úteis, tais como: descrição de produtos, lançamentos de fóruns, satisfação do cliente, etc., para diversas finalidades (Jain *et.al.,* 2013) como melhorar a estratégia de marketing e recomendar os produtos e serviços melhor classificados pelos consumidores. Segundo Linoff e Berry (2001, p.43), os motores de busca, os agentes inteligentes e alguns mecanismos de recomendação implementam *text mining* para ajudar os utilizadores a encontrarem as informações certas no meio da enorme quantidade de dados produzida pela *web*. Alguns motores de busca ou de recomendação de páginas utilizam esta tecnologia para ajudar os utilizadores a encontrar o que procuram na *Internet*. Com isso será possível responder a questões como, por exemplo, onde se encontram os melhores restaurantes de uma dada região.

Com esta ferramenta é possível extrair comentários de clientes de diferentes *sites* para descobrir o que eles pensam e determinar o sentimento subjacente a estes comentários. Este tipo de tarefa já não é possível de ser realizada recorrendo apenas às técnicas tradicionais de *data mining* (Liu, 2012) e em complemento são implementadas técnicas de análise de sentimentos que permitem determinar o sentimento dos textos extraídos.

No início da sua exploração, as aplicações de *text mining* usavam apenas o modelo *bag-of-words* para estruturar os dados, tentando classificá-los com base em duas ou mais classes pré-determinadas ou agrupá-los de forma natural (Sharda *et.al.,*2014, pp.233). Neste tipo de modelo, o texto é representado como um conjunto de palavras, ignorando a gramática e a ordem em que estas aparecem no texto. Mais tarde surgiram alguns estudos que alertaram para o facto de o método *bag-of-words* não ser capaz de produzir um conteúdo suficientemente bom de informação para sustentar a aplicação das tarefas de *text mining,* tais como classificação, associação, *clustering*, etc., pois, naturalmente, os humanos não usam palavras sem uma certa ordem ou estrutura, estas são usadas em frases que possuem estruturas semânticas e sintáticas. Assim sendo, as técnicas automáticas, como o *text mining*, precisam de atingir um nível para além da interpretação do *bag-of-words* e incorporar mais estruturas semânticas nas suas operações. Com isso, a atual tendência do *text mining* está voltada para a inclusão de recursos e de técnicas mais avançadas, tais como o processamento de linguagem natural (PLN) (Sharda *et.al.,* 2014, pp.233) no processamento e análise de dados.

Os dados a utilizar em *text mining* devem passar por várias transformações, na fase de pré-processamento, a fim de construir um bom *dataset* que será utilizado como *input* na construção do modelo. Esta tarefa é levada a cabo pelo Processamento de Linguagem Natural (Miller, 2005; Provost & Fawcett, 2013, pp.253; Sharda *et.al.,* 2014). O processo de *text mining* será dividido em duas grandes fases: a estruturação dos dados (Miller, 2005, pp. 104), seguida da extração de informação e conhecimento através das ferramentas e técnicas de *data mining* (Sharda *et.al.,* 2014, pp.230). A aplicação de PLN envolve análise não estruturada ou linguagem natural gramaticalmente estruturada, criando expressões regulares que serão fáceis de analisar pelo computador (Deepthi.V & Rekha, 2014; Miller, 2005).

2.2.1 Processamento de Linguagem Natural (PLN)

Considerado um componente muito importante do *text mining* que detém as principais tarefas executadas na fase de pré-processamento, possibilitando um primeiro nível de estruturação dos dados*,* o PLN é um subcampo da inteligência artificial e linguística computacional, constituído por um conjunto de técnicas teórico-computacionais que analisam e representam dados textuais com o objetivo de compreender a linguagem humana natural, tornando-os fáceis de serem manipulados pelos programas computacionais. Este tem como objetivo ir além da manipulação de texto orientado à sintaxe para uma verdadeira compreensão e processamento da linguagem natural, que considera contexto e restrições gramaticais e semânticas, aproximando-se o mais possível da linguagem humana (Sharda *et.al.,* 2014, pp.234).

O processo PLN contempla vários níveis de conceitos linguísticos e desafios, tais como o morfológico, que lida com tratamento das palavras, o léxico, que se refere à análise do significado das palavras e do *part-of-speech*, o sintático, que trabalha a gramática e a estrutura das frases, o fonético, que lida com a pronúncia, o semântico, que traduz o significado das palavras e frases, o discurso, que lida com a estrutura de diferentes tipos de texto e, por fim, o pragmático, que introduz o conhecimento presente nas pessoas (Feldman, 1999). Mas nem todas estas áreas são aplicadas ao *text mining*.

Um dos conceitos a ser utilizado no presente projeto é o *part-of-speech tagging*. Este é considerado um processo muito difícil e complexo, uma vez que tenta marcar os termos com uma característica gramatical correspondente no texto, tais como nomes, verbos, adjetivos, advérbios, pronomes, etc., isso porque o *part-of-speech* não depende apenas da definição do termo, mas também do contexto em que este é usado. Este será utilizado no projeto para a identificação das entidades representadas no documento a analisar. Segundo a terminologia básica de *text mining,* um documento é uma peça de texto, quer seja grande ou pequeno, que pode ser uma simples frase ou 100 páginas de um relatório. Um documento é composto por *tokens* ou termos individuais, sendo que estes figuram como palavras. Um conjunto de documentos é denominado por *corpus (*do latim que significa "corpo") e o seu plural é *corpora* (Provost & Fawcett, 2013, p.253; Sharda *et.al.,* 2014, pp.234).

Uma área proeminente que já começou a coletar os resultados e benefícios do PLN é o CRM que, em termos gerais, tem como principal objetivo a maximização do valor do cliente através de uma melhor compreensão e resposta eficaz às suas necessidades reais e percecionadas (Sharda *et.al.,* 2014, pp.234). Um sector muito importante do CRM, onde o PLN está a ter um impacto significativo, é na análise de sentimentos. A análise de sentimentos é uma técnica usada para detetar opiniões favoráveis e não favoráveis relativamente a produtos e serviços usando um grande número de fontes de dados textuais, como exemplo, o *feedback* dos clientes sob a forma de publicações no mundo virtual (redes sociais, fóruns, blogues, etc.). Tradicionalmente, o *text mining* tem sido utilizado para executar a tarefa de análise de sentimentos sobre os dados dos *social media,* utilizando classificação de textos para distinguir os comentários que são positivos, negativos e neutros (Thiel *et.al.,* 2012; Asur & Huberman, 2013). Utilizar os dados públicos disponibilizados *online* para executar a análise de sentimento*s* reduz enormemente os custos, o esforço e o tempo necessário para gerir pesquisas de opinião pública em larga escala e questionários (Bollen *et.al.,* 2009).

2.2.2 Análise de sentimentos

Durante a última década, tem havido um interesse crescente na área do PLN, o principal aspeto da análise de sentimento. As pesquisas variam de classificação de nível de documento (Pang & Lee, 2008) a estudos de polaridade das palavras e frases (Liu, 2012; Sharda *et.al.,* 2014, pp.253).

Com várias nomenclaturas como *emotional polarity analysis*, *review mining, subjectivity analysis, opinion mining* e *appraisal extraction* (Liu, 2012; Sharda *et.al.,* 2014, pp.253), a análise de sentimentos é considerada um processo que deteta automaticamente o conteúdo emocional ou opinativo presente num texto e determina a sua polaridade (Paltoglou & Thelwall, 2012). A polaridade

dos sentimentos é uma característica particular do texto, normalmente dicotomizado em positivo e negativo, ou por um intervalo de valores, onde por exemplo com valores entre]0, 1] o texto é positivo e com valores entre [-1 , 0[, o texto é negativo.

É difícil encontrar na literatura uma definição de análise de sentimentos que seja coerente e comummente aceite pela maioria da comunidade científica. Existem muitas definições em que muitas vezes esta é ligada ou confundida com outros termos como *belief, view, opinion* e *conviction* (Sharda *et.al.,* 2014, pp.253). Mas de entre as muitas definições existentes, encontram-se algumas que vão mais ou menos de encontro com a análise de dados e descoberta de conhecimento que fazem parte do âmbito do presente projeto. Por exemplo, Mostafa (2013) considera a análise de sentimentos uma técnica de descoberta de conhecimento automática que visa encontrar padrões escondidos nos inúmeros comentários textuais. E acrescenta que, para se poder calcular o sentimento do texto obtido, é necessário compará-lo com um léxico ou um dicionário que determina a força do sentimento. Ainda outros autores defendem que este é um procedimento popular do *text mining* que permite ao utilizador final descobrir rapidamente publicações com conteúdo altamente emotivo (Pang & Lee, 2008; Abrahams *et.al.,* 2012).

O sentimento tem algumas propriedades únicas que o diferencia de outros conceitos que podem ser identificados no texto. Normalmente o que se pretende é agrupar o texto envolvendo tópicos e respetivas taxonomias que depois serão classificados de acordo com o tipo de polaridade predominante e respetivo valor. A classificação de sentimentos geralmente lida com duas classes: positivo *versus* negativo, e com o intervalo de polaridades (Sharda *et.al.,* 2014, pp.253; Prabowo & Thelwall, 2009) ou até mesmo com um intervalo de força de opinião (Pang & Lee, 2008). O súbito aumento de interesses e atividades na área de análise de sentimentos à volta da extração automática de opiniões, sentimentos e subjetividade de texto tem criado oportunidades e ameaças para as empresas e indivíduos. Aqueles que o adotam e se aproveitam dele obterão muitos benefícios, pois cada opinião colocada na *Internet* por um indivíduo ou empresa terá conotações positivas ou negativas que podem ser extraídas por outros para diversos fins, sendo a maioria para fins comerciais.

No que respeita ao ambiente empresarial, especialmente na área de marketing e CRM, a análise de sentimentos procura detetar opiniões favoráveis ou desfavoráveis sobre produtos e serviços específicos usando uma grande quantidade de dados textuais virtuais recolhidos dos *social media,* como as redes sociais, *sites* de recomendações, fóruns, blogues, etc. O sentimento que aparece no texto pode ser caracterizado como **explícito**, onde a frase expressa diretamente uma opinião positiva ou negativa, ou **implícito,** onde a frase implica um parecer positivo ou negativo (Liu, 2008, p.421). A maioria dos trabalhos inicialmente realizados nesta área tem foco no primeiro tipo de sentimento por este ser mais fácil de analisar. Atualmente a tendência é implementar métodos analíticos considerando as duas características (Liu, 2008, p.426).

A maioria das pesquisas tem-se centrado em torno de comentários sobre produtos com o objetivo de prever se um consumidor recomenda o produto ou não, com base no conteúdo dos textos. Regra geral, é relativamente fácil detetar se um comentário extraído de um *site* é positivo ou negativo, simplesmente pela extração da *metadata* específica que acompanha o comentário, tal como o número

de estrelas ou os polegares para cima ou para baixo (Paltoglou & Thelwall, 2012). Assim sendo, a análise de sentimentos é uma tarefa de classificação e representa o estudo computacional de sentimentos, subjetividades, avaliações e emoções expressas no texto. As empresas muitas vezes gastam enormes quantias de dinheiro contratando especialistas que tentam encontrar opiniões de consumidores por meio de pesquisas e de grupos específicos.

Devido à complexidade do problema, manifesta nos conceitos subjacentes, expressões em texto e contexto em que o texto é expresso, não há nenhum processo padronizado prontamente disponível para conduzir a análise de sentimentos (Sharda *et.al.,* 2014, pp.258). Na literatura encontram-se muitos trabalhos direcionados para avaliar a polaridade do sentimento ao nível do documento (Pang & Lee, 2008; Sun *et.al.,* 2014). A maioria das aplicações de análise de sentimentos pode ser classificada em quatro categorias distintas: comentários sobre produtos, comentários sobre filmes, extração de orientações políticas e previsão do mercado de ações (Mostafa, 2013). Neste grande grupo destacam-se alguns trabalhos com resultado de sucesso na construção de modelos de análise de sentimentos e análise preditiva recorrendo às técnicas de *text mining,* como é o caso do trabalho realizado por Thiel *et.al.,* (2012). Nele, os autores aplicaram uma combinação das técnicas de *text mining*, análise de sentimentos e *network analysis* sobre os dados do *site* "Slashodot", utilizando a ferramenta KNIME. Outros, como Asur e Huberman (2013) analisaram os *tweets* do Twitter e construíram um modelo preditivo para prever com precisão as receitas de bilheteira na semana de estreia de alguns filmes. Estes tinham como objetivo observar se o conhecimento extraído dos *tweets* poderia permitir uma previsão razoável do futuro no mundo real. Jansen *et.al.,* (2009) também se debruçaram sobre os *tweets* a fim de examinarem um mecanismo para a transmissão WOM relacionada com marcas e produtos específicos, enquanto examinavam a estrutura das publicações e a mudança de sentimentos.

No que respeita à aplicação destas técnicas na área do turismo, não foram encontrados muitos suportes na literatura. As primeiras publicações feitas na área surgiram em 2007 onde três das quais abordaram a influência inesperada dos *social media* nas empresas e indústria do turismo (Zeng, Gerritsen, 2014), indicando que tanto as empresas como a indústria estavam a perder controlo sobre o que estava a ser escrito *online* em relação às mesmas (Dwivedi *et.al.,* 2007), e que a indústria seria confrontada com as consequências, caso os comentários *online* não fossem devidamente geridos, pois os blogues não têm apenas impactos positivos, mas também negativos (Thevenot, 2007). Os autores concluíram que o crescimento e o impacto dos *social media* no turismo e na hospitalidade não deveria ser ignorada. Além disso, os *social media* necessitam de ser constantemente monitorizados pelo turismo com respostas e interações (Grant-Braham, 2007) em tempo real.

Blair-Goldensohn *et.al.,* (2008) apresentaram um trabalho onde utilizaram os dados do *Google Maps* como *input* na construção da análise de sentimentos, de forma a analisar os sentimentos dos consumidores em relação a hotéis, lojas e restaurantes, determinando o valor das polaridades (positivo / negativo). O sistema desenvolvido foi capaz de resumir o sentimento em relação a diferentes aspetos dos serviços prestados, como preço e ambiente. Pekar e Ou (2008) fizeram uso das técnicas de análise de sentimentos para avaliar 268 comentários de clientes de grandes hotéis publicados no *site*

epinions.com. Os autores utilizaram recursos como comida, serviço de quartos, instalações e preço para analisar automaticamente os sentimentos expressos para cada um deles.

Mais na vertente do estudo do presente projeto surge a investigação de Bakhshi *et.al.,* (2015) que também estudou a Yelp para perceber quais os componentes de um comentário de alta qualidade definida pelos consumidores e maximizar a qualidade dos comentários dentro da comunidade. Analisaram 230.000 comentários os respetivos votos (*useful, funny* e *cool*) e mais componentes que constituem um comentário publicado. Os autores descobriram que os membros mais ativos e regulares da comunidade são os que mais contribuem para a boa qualidade dos comentários e que os comentários mais longos têm maior probabilidade de serem populares, recebendo estrelas e votos e que os comentários considerados "useful" tendem a ser os primeiros a serem vistos. A contribuição deles passa pelo incentivo aos membros da comunidade a publicarem comentários com mais qualidade e eficácia.

As ferramentas aplicadas neste tipo de processos utilizam predominantemente técnicas baseadas em dicionário e de aprendizagem automática[5], ou seja, utilizam dados encontrados em recursos lexicográficos para atribuir sentimentos para um grande número de palavras (Lawrence, 2014). Juntamente com um conjunto de algoritmos desconhecidos, essas ferramentas determinam a polaridade de um determinado documento, classificando como sendo positivo, negativo ou neutro. No entanto, a maioria das ferramentas de análise de sentimento oferecem uma classificação avançada dos sentimentos que vai "além da polaridade". Por exemplo, algumas ferramentas analisam os estados emocionais, tais como triste, feliz ou com raiva, enquanto outras determinam o sentimento de uma frase com pontuação específica de sentimento.

Abeywardena (2014) apresentou um estudo onde explorou a opinião pública sobre os termos "MOOC" e "OER" com base nos *tweets* com períodos de 6 e 12 meses, respetivamente. Utilizou *text mining* para analisar os *tweets* e empregou o *software* Semantria para executar a análise de sentimentos, e com isso compreender como as perceções públicas vão mudando durante o período. A construção de um modelo de análise de sentimentos está fora do âmbito deste projeto, onde apenas será utilizado o modelo incorporado no Semantria, perguntando qual o sentimento dos comentários presentes na amostra a analisar.

Uma ferramenta de análise de sentimentos habilmente implementada apoia essas empresas a economizar dinheiro e a antecipar o mercado. Utilizando como exemplo um comentário de um *site* de recomendações (Figura 3), a ferramenta de análise de sentimentos a utilizar no projeto (Semantria), identifica qual o sentimento dominante do comentário.

[5] Tradução adoptada para o termo "Machine Learning".

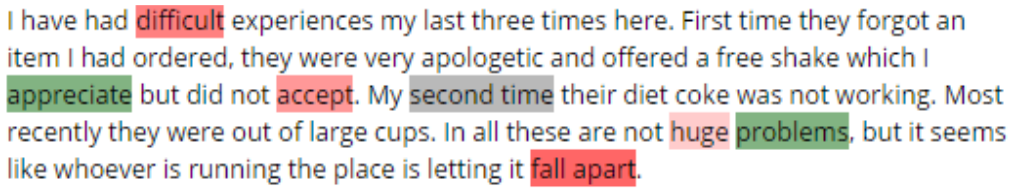

Figura 3 – Exemplo de interpretação de textos com base na análise de sentimentos

Neste exemplo, o consumidor expressa uma opinião claramente negativa da experiência que teve da sua ida a um restaurante. Os termos "difficult experiences", "not accept" e "fall apart" foram determinantes para a classificação do documento onde as frases positivas "appreciate" e "not huge problems" não foram suficientes para inverter a classificação. Na execução da análise de sentimentos, quanto mais avançado o léxico, mais detalhados podem ser a análise e os resultados (Thiel *et.al.,* 2012). Cada frase de um documento é único, composto por nome, verbo, adjetivo, proposições, etc., cada nome representa uma entidade, e num documento pode-se analisar o sentimento do documento num todo e de cada entidade em particular. Os sentimentos são reforçados ou enfraquecidos pelos adjetivos, que podem ser positivos ou negativos. Analisar os sentimentos não é tarefa fácil, pois as frases muitas vezes são ambíguas ou irónicas, o que muitas vezes não produz o resultado pretendido.

As técnicas utilizadas na classificação de sentimentos são principalmente baseadas em métodos heurísticos e em aprendizagem automática (Wang *et.al.,* 2014). Os métodos heurísticos geralmente empregam léxicos pré-definidos e regras de cálculo com base no número total de sentimentos positivos ou negativos (Pang & Lee, 2008), enquanto a abordagem de aprendizagem automática tem recebido atenção significativa para a classificação de sentimentos devido ao seu desempenho na classificação predominante (Pang & Lee, 2008). Através da construção do modelo preditivo, com recurso ao conjunto de dados marcados como dados de treino, pode ser possível modelar mais recursos e adaptar-se às mudanças dos *inputs* mais robustamente, dando-lhes uma vantagem em comparação com os métodos heurísticos (Sun *et.al.,* 2014).

À semelhança de Bakhshi *et.al.,* (2015), o presente projeto também estudará os comentários da plataforma Yelp analisando 14.000 comentários apenas com foco nos seus votos "useful", onde agrupará os comentários em tópicos diferentes e depois analisará os votos e sentimentos de cada um dos tópicos. E no fim, com base no voto "useful" dos comentários dados pelos consumidores será construído um modelo capaz de prever quais os principais termos para determinar a utilidade de comentários futuros.

3 Metodologia

Como referido no Capítulo 1, a metodologia a seguir na etapa do desenvolvimento do presente projeto é o CRISP-DM (*Cross Industry Standard Process for Data Mining*), por ser uma metodologia *standard* aplicada à extração de conhecimento dos dados (Piatetsky, 2014; Sharda *et.al.,* 2014, p.244). Esta encontra-se dividida em seis fases, tratando-se de um processo interativo e iterativo, não sequencial, em que as diferentes fases podem ser executadas mais de uma vez dependendo dos resultados obtidos nas fases seguintes (CRISP-DM, 2000; Larose, 2006). A Figura 4 apresenta a metodologia seguida no presente projeto adaptado do CRISP-DM, contemplando as principais tarefas desenvolvidas no processo de *text mining* defendidas por Sharda *et.al.* (2014, p.244-252).

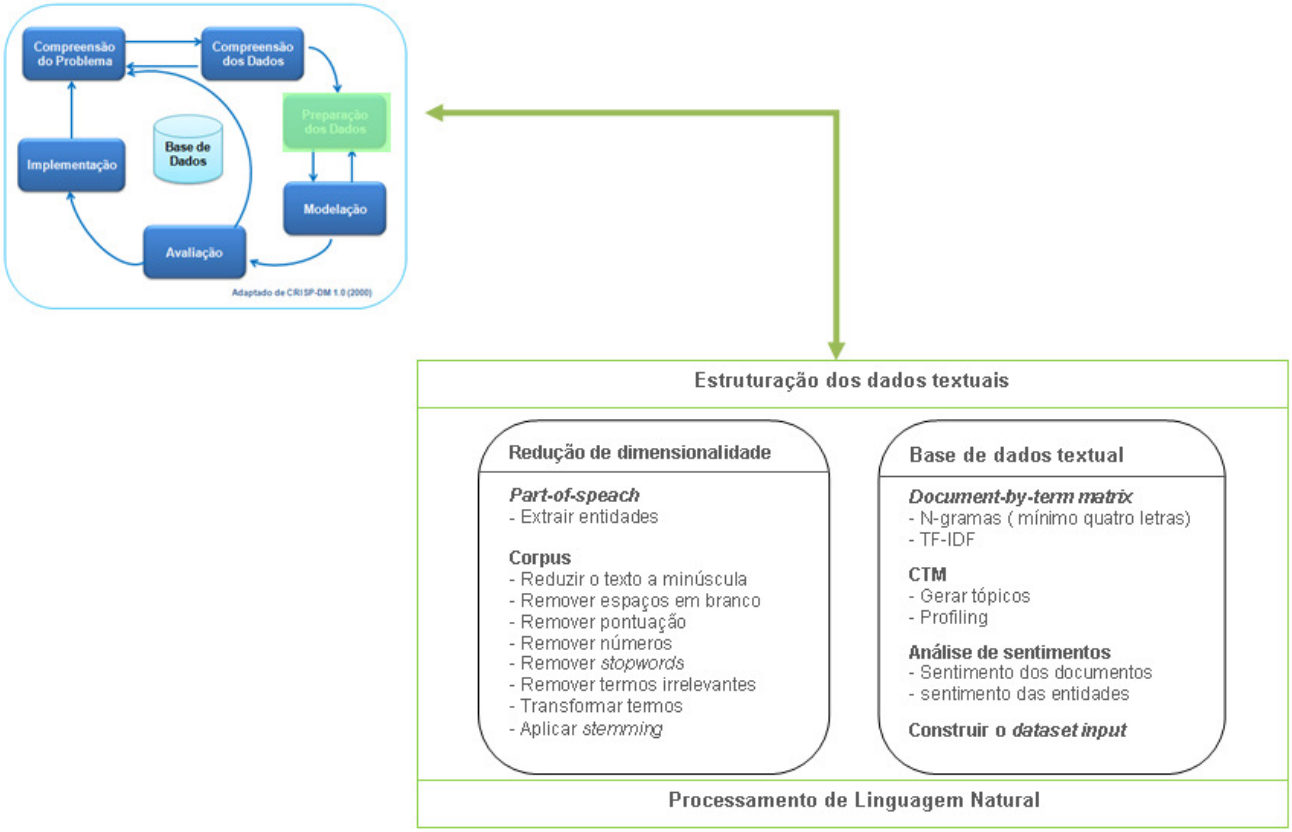

Figura 4 – Metodologia

Fontes: Adaptado de CRISP-DM (2000); Miller (2005); Bollen *et.al.* (2009); Sharda *et.al.* (2014)

A primeira fase da metodologia refere-se à Compreensão do Problema e à sua contextualização, onde é definido o problema e os objetivos propostos. Na fase seguinte é feita a compreensão dos dados recolhidos, incluindo uma avaliação prévia da qualidade dos dados. Segue-se a fase dedicada à Preparação de Dados, sendo esta considerada uma fase essencial para o desenvolvimento do projeto. Como se pode observar na Figura 4, nesta fase recorre-se ao *software* R para a execução das técnicas de *text mining.* Entre elas encontra-se a representação do *Bag-Of-Words*,

o cálculo do TF-IDF, os *N-Grams*, o *Stemming*, o *Named Entity Extraction* e o *Topic models* (Provost & Fawcett, 2013, p.251), que culminam no esforço para a estruturação dos dados. Uma vez estruturadas serão então, aplicadas técnicas de classificação para tentar encontrar o melhor modelo de classificação dos comentários de acordo com os votos (*useful).* Mas*,* ainda antes da criação dos modelos será feita a análise de sentimentos com o auxílio do *software* Semantra onde será determinada a polaridade dos sentimentos (positivo, negativo e neutro) de cada comentário, tópico e entidade.

Na fase de Modelação são aplicadas diversas técnicas de aprendizagem e algoritmos que permitem obter modelos alternativos recorrendo-se ao *software* Modeler. Durante esta fase é muito frequente ter de voltar a efetuar atividades de pré-processamento (Piatetsky, 2015), uma vez que podem ser identificadas variáveis com pouca importância para o modelo.

Os resultados obtidos são avaliados com base em métodos de validação ou teste (fase de Avaliação), ou seja, estes são confrontados com conhecimento prévio e é feita a revisão dos passos efetuados na construção do modelo, a fim de verificar se o resultado do modelo se encontra alinhado com os objetivos propostos.

A fase de Implementação é reservada à apresentação dos resultados obtidos da modelação. Tal como em todas as metodologias, o CRISP-DM também não garante resultados, mas permite disciplinar o processo de desenvolvimento de um modelo alinhando-o com os objetivos do projeto (Piatetsky, 2015).

A escolha dos *softwares* está relacionada com o facto de estes serem os mais utilizados, com sucesso, na área do *text mining*. O R lidera o *ranking* das ferramentas mais utilizadas no desenvolvimento de projetos de *text mining* (Piatetsky, 2015), enquanto o Semantria pertence à Lexalytics, líder na indústria de transformação de textos e análise de sentimentos. A IBM SPSS Modeler é uma poderosa ferramenta de análise preditiva que oferece um vasto leque de algoritmos e técnicas que permitem obter os melhores resultados de análise.

3.1 Compreensão do problema

A quantidade de dados existentes em formato eletrónico não estruturado tem vindo a crescer ao longo dos tempos e outrora a sua importância foi ignorada, perdendo-se assim um grande contributo para o negócio das organizações em geral. Como anteriormente referido, a capacidade do ser humano para processar informação é limitada. Ler, compreender e sintetizar gigabytes de textos diariamente é uma tarefa com uma densidade de informação impossível de interpretar. Isto levou a que muitas informações ficassem por encontrar e como consequência muitas oportunidades fossem provavelmente desperdiçadas.

Para tentar combater este desperdício, vários investigadores vêm ganhando interesse na exploração de estratégias para a gestão de informação a fim de estabelecer uma certa ordem no crescente volume de dados não estruturados. Este fenómeno deve-se ao surgimento da *Web 2.0* e dos *social media* que também não param de crescer proporcionando nos últimos anos um aumento significativo de comentários *online* que estão cada vez mais acessíveis aos consumidores, ajudando-

os e influenciando-os nas suas tomadas de decisão. Os comentários ajudam, por um lado, as empresas a recolher informações quanto à perceção dos consumidores em relação aos bens e serviços. Por outro lado, ajudam e influenciam os consumidores a centrarem a sua atenção nas recomendações que poderão estar mais alinhadas com o preenchimento das suas necessidades, recomendações essas que são o resultado da filtragem de uma grande quantidade de informação que poderá não responder a essas necessidades. No entanto, neste ambiente de informação complexa pode-se encontrar muita informação que carece de confiabilidade, fidedignidade e utilidade (Hajas *et.al.,* 2014) dos comentários encontrados, pelo que é importante questionar se todos estes comentários serão úteis para os consumidores.

Hoje em dia, todos os consumidores e empresas estão dependentes da resposta a esta questão, mas para a indústria turística esta resposta é essencial para continuar a crescer, pois os turistas necessitam de garantias dos bens e serviços que estão prestes a adquirir e as empresas turísticas necessitam de testemunhos de clientes satisfeitos para aumentarem a sua confiança e credibilidade.

Para ajudar nesta questão, surgiram vários s*ites* de comentários *online* entre os quais as plataformas de recomendação que se preocupam em, para além de disponibilizar os comentários atualizados, tentar garantir que estes sejam de facto úteis para a tomada de decisão. Mas infelizmente nem sempre isso acontece e mediante a abundância de conteúdos gerados pelos consumidores, para cada opinião interessante ou comentário útil, há conteúdos e opiniões que são inúteis, subjetivos ou enganosos. E filtrar grandes quantidades de comentários para identificar uma informação que seja útil é um processo tedioso e propensos os erros (Bakhshi *et.al.,* 2015). Um exemplo de mecanismo de avaliação dos comentários são os votos e um dos *sites* que o disponibiliza é a plataforma Yelp.com. Esta permite que os consumidores classifiquem um comentário como *funny, cool* e/ou *useful*. Compreender estes votos pode ajudar as plataformas de recomendação a desenhar mecanismos para melhorar a qualidade das publicações (Bakhshi *et.al.,* 2015).

No presente trabalho, pretende-se responder a questão de investigação: Quais os termos que determinam a utilidade de um dado comentário publicado pelo consumidor? A maior preocupação do projeto passa por distinguir, no meio de uma grande quantidade de comentários, os que são úteis quer para os consumidores como para as empresas, auxiliando-os nas suas tomadas de decisões futuras. Mais antes será respondida a questão: Qual o sentimento que caracteriza cada um dos comentários?

3.2 Compreensão dos dados

Os dados utilizados foram obtidos através da plataforma Yelp na parte do Yelp Dataset Challenge[6]. Estes formam uma coleção de ficheiros .json divididos entre *user*, *business, review, check.in* e *tip*. Mas como o objetivo do presente projeto é analisar os comentários não estruturados, os dados obtidos foram

[6] https://www.yelp.com/dataset_challenge

todos do *dataset Yelp_academic_dataset_review.json* onde se encontra o conteúdo dos comentários e os respetivos votos *useful*. Este encontra-se estruturado de acordo com a Figura 5.

```
{
    'type': 'review',
    'business_id': (encrypted business id),
    'user_id': (encrypted user id),
    'stars': (star rating, rounded to half-stars),
    'text': (review text),
    'date': (date, formatted like '2012-03-14'),
    'votes': {(vote type): (count)},
}
```

Figura 5 - Estrutura do dataset yelp_academic_dataset_review.json

O *dataset* é constituído por um total de 1.569.264 comentários sobre as mais diversas empresas em redor das universidades de Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas e Madison nos Estados Unidos. Cada observação do *dataset* é caracterizado por 10 variáveis: *user_id, review_id, stars, date, text, type, business* e *votes,* sendo o *votes* dividido em: *funny, useful* e *cool.* O *user_id* serve para associar o comentário a outros do mesmo consumidor e o *business_id* para associar o comentário a outros da mesma empresa.

Por ser um *dataset* muito grande e por não haver recursos suficientes para analisar a totalidade dos dados, foi necessário extrair uma amostra aleatória com informação passível de ser analisada em tempo útil pelos algoritmos apresentados com os recursos tecnológicos disponíveis e pelo *software* Semantria, que na altura do desenvolvimento do projeto oferecia uma capacidade máxima para *trial* de 15.000 registos. O objetivo do projeto passa, então, pela elaboração do relatório de sentimentos, construção do *dataset* que será o *input* na construção do modelo e a construção do modelo em si, usando uma amostra de 14.000 registos, que seja capaz de prever a utilidade de um qualquer novo comentário do utilizador, olhando para os seus termos mais frequentes. A Tabela 1 apresenta a descrição de cada uma das variáveis do *dataset* original.

Tabela 1 – Variáveis do *dataset*

Review		
Nome da variável		Descrição
votes	**funny**	Número total de votos *funny*
	useful	Número total de votos *useful*
	cool	Número total de votos *cool*
user_id		Código único que permite identificar cada utilizador
review_id		Código único que permite identificar cada comentário
stars		Número de estrelas que classifica o comentário (1-5)
date		Data em que foi publicado o comentário

Review	
Nome da variável	Descrição
text	O conteúdo do comentário
type	Tipo de objeto *dataset (review)*
business_id	Código único que permite identificar a empresa sobre a qual recai o comentário

A Tabela 2 apresenta alguns valores de análise descritiva de cada uma das variáveis apresentadas na Tabela 1. Nota-se que num total de 14.000 comentários a média dos votos (qualquer um) não atinge sequer os 2 valores. Um valor muito baixo para um máximo acima dos 30. Isto deve-se ao facto de mais de 90% dos comentários apresentarem valores de votos iguais ou muito próximos de zero.

Tabela 2 – Análise descritiva das variáveis

Varáveis	Mín.	Máx.	Média	Desvio padrão	n
votes_funny	0	53	0,49	1,64	-
votes_useful	0	38	1,09	2,19	-
Votes_cool	0	50	0,6	1,71	-
user_id	-	-	-	-	14000
review_id	-	-	-	-	-
stars	1	5	3,75	1,31	-
date	29-09-2006	08-01-2015	-	-	-
text	-	-	-	-	14000
type	-	-	-	-	14000
business_id	-	-	-	-	14000

De entre estas variáveis, foram selecionadas as variáveis *text* e *votes_useful.* O *votes_useful* será a variável objetivo do projeto. Como se pode verificar na Figura 6, esta apresenta uma distribuição completamente enviesada para a direita, onde apenas o zero representa 51% dos votos e 49% dos votos são distribuídos pelos outros números, sendo que o maior número de votos ocorre uma única vez, ou seja, quanto maior o número de votos menor é a sua frequência. O que significa que os elevados números de votos têm menor probabilidade de acontecer.

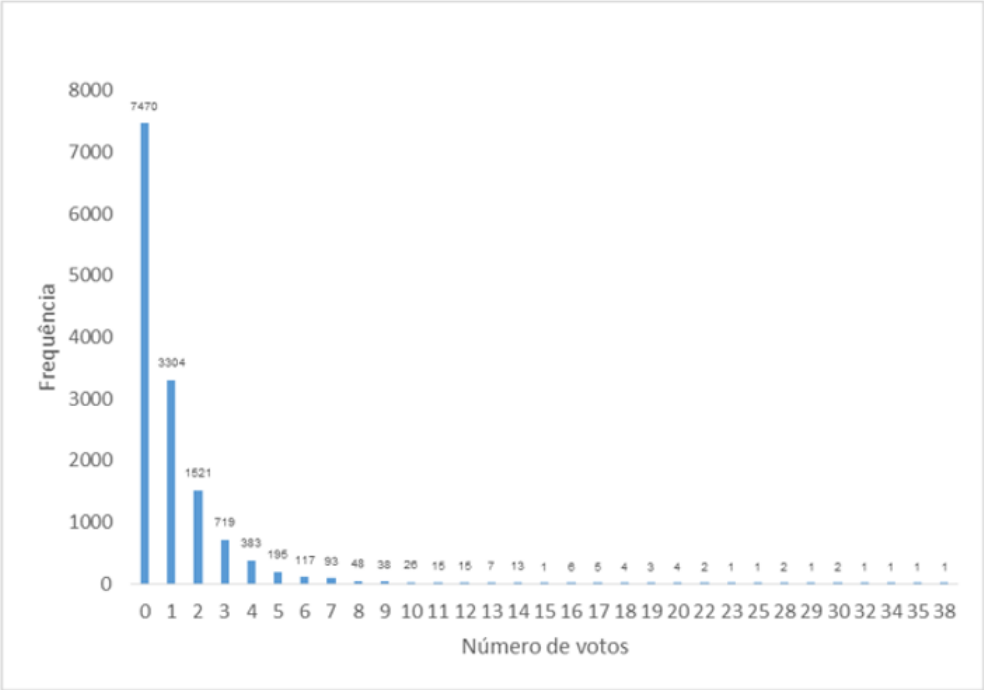

Figura 6 – Distribuição da variável *target* (*votes_useful*)

Para além do *dataset* uma outra informação muito importante também foi obtida. Trata-se das diferentes categorias pelas quais se agrupam os diferentes negócios na Yelp. Estas categorias encontram-se divididas em 22 categorias principais: *Active Life, Art & Entertainment, Automotive, Beauty & Spas, Education, Event Planning & Services, Financial Services, Food, Health & Medical, Home Service, Hotels & Travel, Local Flavor, Local Services, Mass Media, Nightlife, Pets, Professional Services, Public Services & Government, Real Estate, Religious Organizations, Restaurants* e *Shopping*. Para além disso, cada uma destas categorias principais são divididas em subcategorias formando um total de 498 categorias secundárias e algumas delas ainda contêm categorias terciárias num total de 178 categorias. Estas categorias serviram de base para a escolha dos tópicos onde se agrupam os diferentes comentários e foram também uma peça muito importante no processo de *profiling*.

3.3 Preparação dos dados

Considerada uma das fases mais importante, esta tem como principal objetivo estruturar o texto de forma a ser manipulado pelos algoritmos de extração de padrões (Liu, 2008, p.450). Esta etapa normalmente diferencia os processos de *text mining* dos de *data mining* uma vez que num projeto de *data mining* os dados, regra geral, já se encontram estruturados.

O processo teve início com a importação de todo o ficheiro *Yelp_academic_dataset_review.json* no R recorrendo ao *package jsonlite* através da função *stream_in*

(Ooms *et.al.,* 2015). O ficheiro .json foi finalmente convertido para .csv para ser mais facilmente gerido e analisado no R.

Tendo a amostra, para esta primeira fase de pré-processamento de dados foi criado um *dataset* só com a variável *text* e com ele foi construído o *corpus* onde foram aplicadas transformações detalhadas em pormenor por Graham (2014):

- Conversão para minúscula;
- Remoção de números;
- Remoção de pontuação;
- Remoção de *stopwords* (Liu, 2008, p. 199);
- Remoção de mais palavras considerados irrelevantes para a análise e diretamente relacionadas com, por exemplo: yelp, URL e yummy;
- Transformação de alguns termos considerados importantes (Bollen *et.al.,* 2009). Por exemplo, passar todos os bbq para *barbecue*;
- Aplicação do processo de *stemming* (Liu, 2008, p.200) e,
- Remoção de espaços em branco.

A remoção da pontuação permite eliminar todos os sinais de pontuação, as chavetas, as aspas, os asteriscos, etc. Com a sua execução, e uma vez que já se sabe que os comentários muitas vezes são escritos sem nenhuma regra, detetou-se que algumas palavras ficavam coladas umas às outras como demostra a Figura 7. Isto porque depois do carácter ponto não existia espaço antes de começar a frase seguinte. Este facto produzia termos incorretos e muito longos prejudicando assim um pouco a análise. Uma solução encontrada foi a de substituir as pontuações por um espaço em branco em vez de as remover.

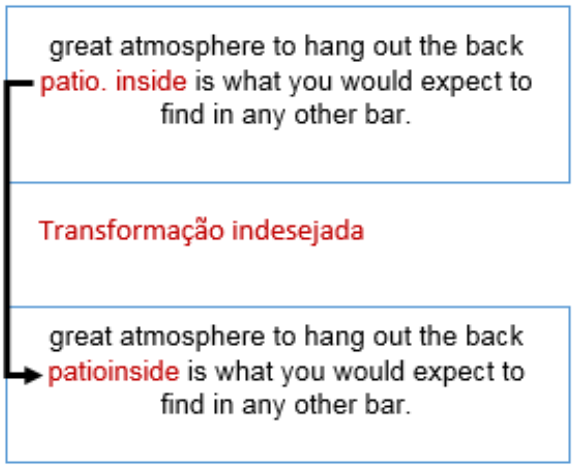

Figura 7 – Exemplo de uma transformação indesejada

A remoção das *stopwords* consiste na eliminação de tudo que sejam verbos auxiliares, artigos, pronomes, preposições, interjeições e mais termos comuns e irrelevantes (Liu, 2008, p.199). Notou-se que por alguma falha no R, nem todas as *stopwords* estavam a ser removidas, pelo que depois da execução do código, procedeu-se à remoção de mais algumas recorrendo a código feito à medida,

mesmo depois da construção da DTM. Isto é um dos processos que obrigou também a voltar atrás e fazer alterações, voltando a correr o código. O processo de *stemming* permite que palavras semelhantes sejam reduzidas aos seus radicais, a fim de não serem identificados como sendo diferentes. O algoritmo de *stemming* elimina os prefixos e sufixos de cada termo. Por exemplo, as palavras *singer* ou *singing* são reduzidos ao seu radical *sing* (Porter, 1980; Liu, 2008, p. 200).

Seguidamente às transformações foi construído o *Document-by-term matrix* (DTM) aceitando *unigramas* e *bigramas* não inferiores a três carateres. Esta é, provavelmente, a forma mais comum de estruturar os dados contidos no *corpus*, onde as linhas representam os documentos, as colunas representam os termos, e na sua interceção se encontra o valor da frequência absoluta do termo no documento (Feinerer *et.al.,* 2008).

Numa primeira análise sobre a DTM detetou-se que a transformação deu origem a um número muito elevado de termos, gerando um elevado grau de dispersão contendo termos muito extensos (Figura 8). Estes resultados são problemáticos uma vez que indica a existência de termos que raramente são mencionados nos comentários e que podem ser pouco relevantes para a análise. Para obter uma matriz mais consistente procedeu-se à alteração de alguns parâmetros tais como: (1) no "tokenize" foi indicado um limite de ngramas de 1 até 3 bigramas no máximo, (2) estabeleceu-se que o tamanho mínimo do termo *(minWordLength*) a ser considerar seria 4, (3) o mínimo de frequência do termo no documento (*minDocFreq*) seria 2 e (4) voltou-se a executar os métodos *stemming*, *stopwords* e *removeNumber* que já tinham sido aplicados sobre o *corpus*.

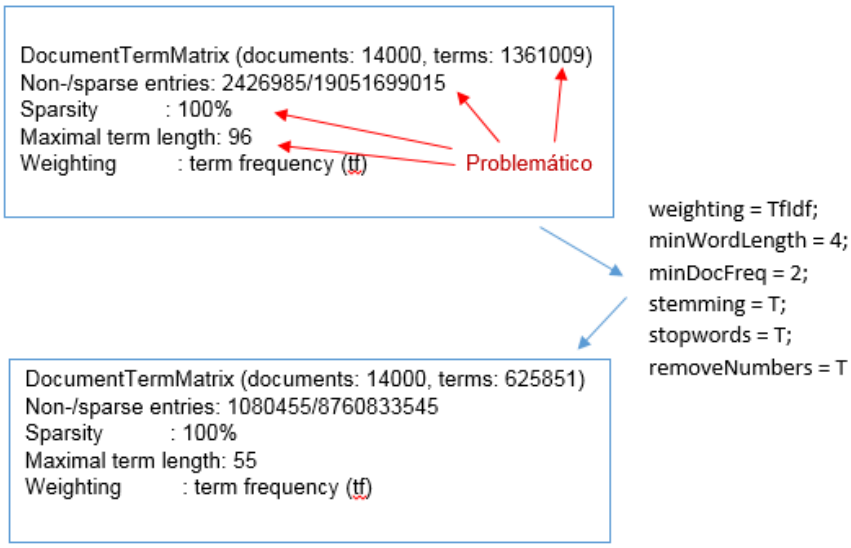

Figura 8 – Melhorias aplicadas ao DTM

Como se pode verificar na Figura 8 a DTM revelou que para os 14.000 documentos foram detetados 625.851 termos com 100% de dispersão.

Geralmente, o desempenho dos algoritmos de reconhecimento de padrões é muito prejudicado com base de dados dispersas e de alta dimensionalidade. A grande dimensionalidade provoca um alto custo computacional, tornando a execução dos algoritmos muito lenta e até inviável em vários casos

(Feinerer *et.al.,* 2008). Perante este cenário, é fundamental a aplicação de técnicas de redução da dimensionalidade, ou do número de termos, para melhorar a eficácia e eficiência dos algoritmos de reconhecimento de padrões. O objetivo é utilizar apenas os termos mais relevantes para representar os documentos no domínio do problema (Feinerer *et.al.,* 2008).

Isto levou a que fosse aplicada uma medida para reduzir a dispersão. A medida adotada foi o *Term Frequency Inverse Document Frequency* (TF-IDF) (Blei & Lafferty, 2007) que combina a frequência do termo com o inverso da frequência do documento, favorecendo termos com alta frequência de documento e que apresentam uma distribuição não uniforme ao longo do *dataset*. É usado para eliminar os termos que se repetem bastante num único documento, mas muito pouco nos restantes documentos do *corpora*. Com a aplicação desta técnica apenas os termos com frequência no *corpora* superior à mediana (Grün & Hornik, 2011) continuaram a fazer parte da DTM. Depois deste processo a DTM ficou reduzida a 13.917 documentos para 339.928 termos.

Sobre a DTM procedeu-se a uma análise que permite ver a correlação entre os termos através da função findAssocs (x, terms, corlimit), onde o x deve ser substituído pela DTM, *terms* deve ser o termo em análise e no *corlimit* deve ser indicada a percentagem mínima de correlação que se pretende encontrar. A Tabela 3 apresenta os três termos mais fortemente correlacionados com os 10 termos mais frequentes do *corpora* apresentados pelos seus radicais, resultado da aplicação do *stemming* ao *corpus*.

Tabela 3 - Os termos mais correlacionados com cada um dos 10 termos mais frequentes

Termos	Termos correlacionados	%
Pizza	Pizza pizza	0.6
	Pizza place	0.44
	Crust	0.36
Burger	Burger burger	0.51
	Burger fri	0.41
	Fri burger	0.31
Sandwich	Sandwich sandwich	0.34
	Sandwich place	0.32
Sushi	Roll	0.46
	Sushi place	0.46
	Sushi chef	0.44
Steak		
Roll	Roll roll	0.52
	Sushi	0.46
	Sushi roll	0.4
Breakfast	Breakfast buffet	0.31
	Place breakfast	0.31
Coff	Coff shop	0.41

Termos	Termos correlacionados	%
	Coff coff	0.33
Show	Show show	0.43
	Cirqu	0.34
	Show time	0.34
Car	Car car	0.46
	Work car	0.41
	Car problem car	0.39

Esta tabela mostra que os termos do *corpora,* no geral, não apresentam uma correlação muito forte, ou seja, são raros os casos em que a correlação é superior a 0.5% e mesmo assim os resultados apresentados não são casos comuns, uma vez que as correlações mais fortes são detetadas entre os mesmos termos (exemplo: o termo pizza, está fortemente correlacionado com o termo pizza pizza). A diferença é que o termo da correlação é um ngrama mal formado na construção da DTM e pode-se verificar que há termos que não estão minimamente correlacionados com nenhum dos termos existentes no *corpora (Steak)*. No geral, pode-se dizer que em todo o *corpora* os termos têm uma correlação muito fraca.

Depois de verificadas as correlações foi feita uma análise de frequência dos termos que deu origem ao *wordcloud*. Os primeiros gráficos construídos serviram para detetar os termos que não são relevantes para a análise, porque fazem parte do processo, e termos que são *stopwords* que não foram detetados e removidos anteriormente na limpeza do *corpus*. Para tentar eliminar estas falhas, com o intuito de melhorar a análise, todo o processo atrás foi repetido até este ponto. O resultado é apresentado na Figura 9 contemplando apenas os 68 termos (com *stemming*) mais frequentes (frequência mínima de 200), onde o tamanho das letras é proporcional à frequência de cada termo (os termos mais salientes ocorrem com maior ocorrência).

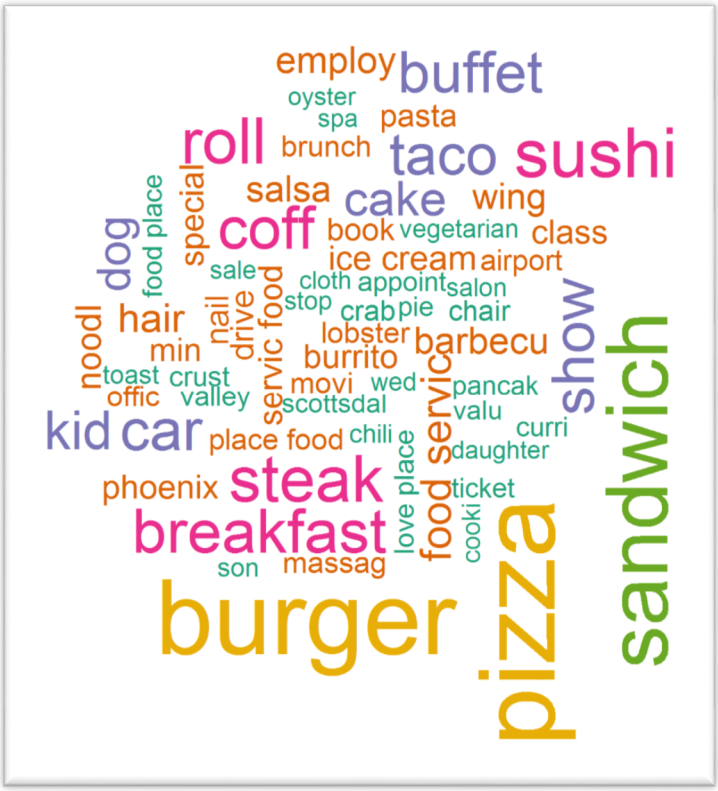

Figura 9 - *Wordcloud* com os radicais dos 68 termos mais frequentes

Apesar de o projeto estar voltado para os comentários relacionados com os vários produtos turísticos, na amostra em análise os comentários predominantes estão, na sua maioria, relacionados com área da restauração abordando os diferentes restaurantes, e realçando os diferentes pratos e serviços prestados. Como resultado, o *wordcloud* manteve o foco nos termos mais relacionados com esta área.

Tendo terminado a análise sobre a DTM, o próximo passo foi agrupar os termos em diferentes tópicos de acordo com a correlação existente entre os termos. No início, os *clusters* na área de *text mining* eram construídos com base em algoritmos de *clustering* tradicionais (Guerreiro, Rita, & Trigueiros, 2015). Entretanto surgiram na literatura modelos que permitem uma análise de pertença *mixed* (um termo pode estar em vários clusters) ao invés da tradicional análise de pertença individual (um termo pode estar apenas em um único cluster) (Blei & Lafferty, 2007; Griffiths & Steyvers, 2004), Um destes modelos é o *topic models* (Guerreiro *et.al.,* 2015), um modelo probabilístico que permite descobrir a estrutura semântica subjacente da coleção de documentos com base na análise Bayesiana (Blei & Lafferty, 2009). Este modelo tem sido aplicado a vários tipos de documentos, incluindo *e-mails*, trabalhos científicos (Blei et al. 2003; Griffiths & Steyvers, 2004) e arquivo de jornais (Wei & Croft 2006). O *Topic Models* permite que os termos pertençam a vários tópicos em simultâneo (Grün & Hornik, 2011). Cada tópico representa uma distribuição de termos onde cada termo tem uma probabilidade diferente de pertencer a esse tópico. Um dos algoritmos de *cluster* de *topic model* mais commumente usado é o *Correlated Topic Model* (CTM) (Blei & Lafferty, 2007). Este algoritmo é baseado no *Latent Dirichlet Allocation* (LDA), onde se assume que os tópicos são independentes (Blei & Lafferty, 2009).

O CTM apresenta claras vantagens sobre os tópicos de *bag-of-word* localizados pelo LDA (Blei & Lafferty, 2007), sendo uma delas a eficácia em termos de *perplexity* relativamente ao LDA (uma medida comummente aplicada a fim de determinar o quão bem o modelo prevê as restantes palavras que irão aparecer num determinado tópico depois de observar uma pequena parte dela) (Guerreiro, Rita, & Trigueiros, 2015). Os autores de CTM compararam ambos os algoritmos (LDA e CTM) e descobriram que o CTM reduz a *perplexity* sobre o LDA pelo menos em 10% (Blei & Lafferty, 2007). O número de *clusters,* para o presente projeto, foi determinado com base na avaliação dos valores da média apresentada no gráfico do *log-likelihood,* onde foi avaliado como estes valores se alteram à medida que iam sendo aumentados os números de *clusters*. O número ideal de *clusters* é determinado quando a variabilidade explicada não se altera significativamente com o incremento de clusters (Guerreiro *et.al.,* 2015).

De forma a testar potenciais modelos CTM com *clusters* que variaram entre os modelos com apenas 2 tópicos e modelos com 60 tópicos, e assim avaliar a sua *log-likelihood* e *perplexity,* foi necessário executar vários modelos com um esforço computacional bastante elevado (72 horas de processamento de máquina). A Figura 10 apresenta o gráfico da média do *log-likelihood* e *perplexity* com a média de 60 prováveis tópicos.

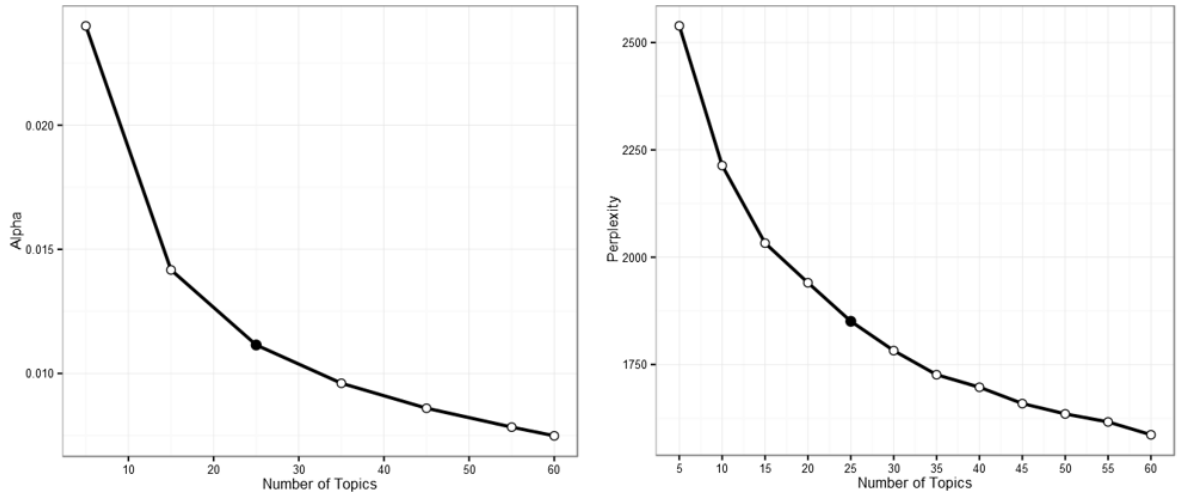

Figura 10 – Valores de Log-likelihood (Alpha) e Perplexity do CTM por número de tópicos

De acordo com o gráfico da Figura 10, o número de tópicos ideais situa-se entre os 15 e os 25. Assim, optou-se por considerar 20 tópicos na tentativa de os comparar com as 22 categorias de negócio apresentados no *site* na Yelp[7].

Apesar de inicialmente se terem construído 22 tópicos com o intuito de replicar as categorias de negócio da Yelp referidas anteriormente, entretanto os tópicos gerados não apresentavam grandes diferenças entre si e acabavam por repetir muito os termos. Desta forma, através de uma análise de

[7] https://www.yelp.com/developers/documentation/v2/all_category_list

sensibilidade optou-se por manter os 20 tópicos, que embora também apresentem algum desequilíbrio, têm menos repetições de termos pelos diferentes tópicos. Uma comparação inicial dos tópicos gerados com as categorias da Yelp sugere que cada um dos tópicos não tem uma correspondência direta com cada uma das categorias apresentadas. A amostra com os tópicos gerados pelo algoritmo de *topic models* representa mais a categoria *Restaurants,* como é o caso dos tópicos *Japanese Restaurant*, *American Restaurant*, *Mexican Restaurant*, *Steak house*, *Pizza restaurant*, *Fast Food Restaurant*, *Buffet*, e *Italian Restaurant*. Refletindo a categoria *Food* estão os tópicos *French Bakery* e *Food*. Relacionados com a categoria *Arts & Entertainment* estão os tópicos *Music Venue*, *Casino* e *Entertainment* e a categoria *Hotel & Travel* está representada pelos tópicos *Hotel* e *Air Travel*.

Os restantes tópicos apresentam uma mistura de categorias pelo que é mais difícil determinar a qual pertencem univocamente. No entanto, uma análise de sensibilidade sugere que seria lógico pertencerem à categoria *Shopping* por ser uma categoria que engloba várias áreas. Se a categoria *Shopping* for pensada como o nome dado aos grandes espaços comerciais que existem, então faz todo o sentido relacioná-la com os tópicos *High Class Place*, *Meeting Place*, *Shopping Center*, *Lounge Area* e *Restaurant and Nails Salon.* Com isso conclui-se que das 22 categorias apenas 4 foram contempladas nos tópicos construídos. Nota-se que há uma mistura de categorias em vários tópicos, o que dificultou o processo de *profiling.* No entanto, com a análise de alguns comentários relacionados com cada um dos tópicos foi possível chegar às nomenclaturas apresentadas na Tabela 4. Esta tabela apresenta os 20 tópicos e os cinco termos mais correlacionados com cada um e as respetivas designações, resultado do processo de *profiling.*

Tabela 4 – Tópicos

1 - Buffet	**2 - American Restaurant**	**3 - Hotel**	**4 - Mexican Restaurant**	**5 - Pizza Restaurant**
941 reviews	856 reviews	786 reviews	760 reviews	760 reviews
Buffet	Burger	Pool	Taco	Pizza
Breakfast	Coffee	Show	Salsa	Crust
Cake	Ice cream	Burger	Steak	Store
Roll	Breakfast	Store	Store	Burger
Pasta	Sandwich	Customer service	Burger	Sandwich
6 - Shopping Center	**7 - Lounge Area**	**8 - Japanese Restaurant**	**9 - Restaurant and Nails Salon**	**10 - Italian Restaurant**
739 reviews	730 reviews	729 reviews	705 reviews	682 reviews
Store	Massage	Sushi	Sandwich	Steak
Taco	Show	Roll	Pizza	Pizza
Sandwich	Pizza	Sushi place	Nail	Sandwich
Nail	Store	Sandwich	Breakfast	Hair
Tire	Office	Sushi bar	Show	Wing

11 - French Bakery	12 - Entertainment	13 - Music Venue	14 - Air Travel	15 - High Class Place
680 reviews	678 reviews	658 reviews	657 reviews	652 reviews
Sandwich	Movie	Show	Flight	Class
Crepe	Pool	Burger	Store	Pizza
Store	Theater	Store	Airport	Burrito
Customer service	Show	Donut	Noodle	Tour
Show	Store	Wing	Show	Noodle
16 - Meeting Place	**17 - Casino**	**18 - Fast Food Restaurant**	**19 - Steak House**	**20 - Food**
645 reviews	644 reviews	630 reviews	627 reviews	353 reviews
Store	Game	Burger	Burger	Food price
Burger	Pizza	Chili	Barbecue	Food service
Buffet	Burger	Sandwich	Coffee	Food food
Mall	Store	Pizza	Sandwich	Food atmosphere
Game	Wing	Coffee	Chop	Folk

Entre as análises efetuadas no processo de *profiling* encontra-se a frequência de cada um dos termos que constituem um determinado tópico no conjunto de documentos e a correlação entre elas. Ou seja, foi verificado se os termos eram referidos juntos no mesmo documento. Em última tentativa verificou-se se nos comentários havia alguma palavra que pudesse englobar todos os termos presentes nos tópicos.

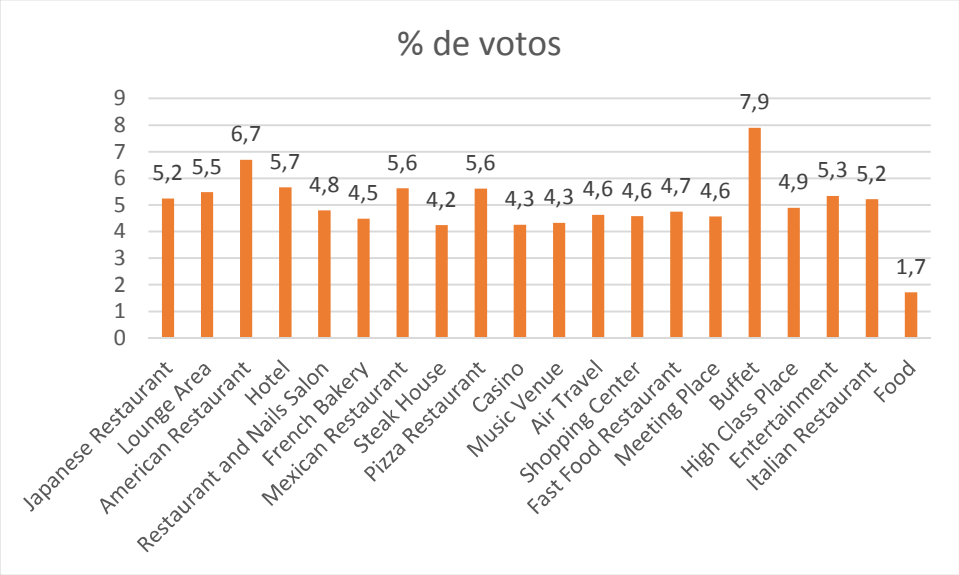

Figura 11 – Percentagem de votos "useful" por tópico

Tendo os tópicos todos definidos, foi feito o levantamento do comentário mais correlacionado com cada um dos tópicos e de seguida fez-se uma soma dos votos que cada um dos comentários

apresenta para determinar o total de votos de cada tópico. A Figura 11 apresenta o resultado desta análise em valores percentuais.

Na relação Comentário-Tópico detetou-se que 88 comentários não estavam correlacionados com nenhum dos tópicos; isto deve-se ao facto de muitos apresentarem caracteres estranhos, sinais de pontuação ou frases pouco frequentes. Tudo isto desapareceu com o tratamento dos dados no R e estes comentários não contaram para a construção dos tópicos. Posto isto, o resultado da análise apresentada na Figura 11 contempla 13.913 comentários e um total de 14.718 votos. E como se pode verificar o tópico "Buffet" lidera com 7,9% do total dos votos seguido do tópico "American Restaurant" com 6,7% dos votos. O tópico "Food" é o que tem menos votos, apresentando apenas 2% do total dos votos. Isto significa que os consumidores consideram que os comentários pertencente aos tópicos "American Restaurant" e "Buffet" são mais úteis dos que os comentários pertencentes ao tópico "Food".

Saindo dos tópicos, o próximo passo é em direção à análise de sentimentos. Aqui uma das principais tarefas a desenvolver foi a execução do *Part-Of-Speech* (POS) *tags*, onde cada comentário foi divido em frases e para cada frase foram identificados os diferentes componentes. O POS *tagging* é uma das técnicas fundamentais do panorama linguístico que consiste na análise sintática básica que tem inúmeras aplicações no PLN (Gimpel, et al., 2011; Chopra & Bangalore, 2012). Existem, na gramática, nove partes do discurso: nome, verbo, artigo, adjetivo, preposição, pronome, advérbio, conjunção e interjeição, cada uma das partes é classificada com uma *tag,* como mostra o exemplo da Figura 12. De acordo com o problema de investigação, vão sendo aplicadas análises sobre diferentes tipos de *tags* (Gimpel, et al., 2011; Chopra & Bangalore, 2012). Para o presente projeto, uma vez que se vai analisar o sentimento dos comentários, o que interessa são as entidades (as *tags* NN, NNP, NNS e NNPS), para que seja possível identificar sentimentos específicos para cada entidade presente no comentário, e não apenas o sentimento do documento.

Figura 12 - Exemplo de um comentário realçando o POS tags

Na preparação para a análise foram adicionados ao Semantria os tópicos construídos com o CTM e as entidades identificadas no POS e que ocorrem pelo menos 200 vezes na DTM para se poder obter os sentimentos individuais dos comentários, tópicos e entidades. Isto porque, por exemplo, pode encontrar-se um documento cujo sentimento seja negativo, mas a entidade referida pode ter um sentimento positivo e assim é mais fácil identificar qual a área que requer mais atenção e melhoria, na ótica do gestor.

O *plug-in* do Semantria no Excel realiza uma análise automática dos sentimentos com base em algoritmos desenvolvidos para extrair o sentimento de um modo semelhante aos seres humanos. A extração de sentimentos de um documento adere os seguintes passos:

1) O documento é dividido em *parts of speech (POS) Tags*,
2) O algoritmo identifica as frases que suportam o sentimento,
3) Inclui uma escala logarítmica de -10 a 10 pontos para cada frase que suporta o sentimento,
4) As pontuações são combinadas para determinar o sentimento geral (*overall*).

Através destas inferências estatísticas, cada comentário é marcado com um valor de sentimento numérico que varia de -2,0 a +2,0 e uma polaridade:

- [-2, -0.05 [→ Negativo
- [-0.05, 0.22 [→ Neutro
- [0.22, 2] → Positivo

Para a classificação das entidades, categorias e mais componentes disponibilizados pelo Semantria, o intervalo varia entre -10 e 10. Nota-se que, quanto maior o valor do sentimento maior é a positividade (Abeywardena, 2014).

Lawrence (2014) defende que o Semantria funciona como uma *black box,* ou seja, um sistema onde apenas é possível visualizar o *input* e o *output* como mostra a Figura 13. As funcionalidades internas e características de transformação do *input* em *output* não são conhecidas. O que se sabe é que a *black box* consiste num conjunto de algoritmos e léxicos que permitem identificar frases portadoras dos sentimentos (Lawrence, 2014).

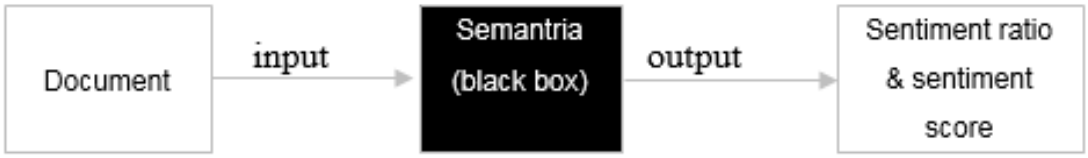

Figura 13 - Modelo de Análise de Sentimentos
Fonte: Adaptado de Lawrence (2014)

Como resultado da análise de sentimentos obteve-se o *overall* de sentimentos para cada comentário e os sentimentos de cada entidade e categoria relacionados com os comentários. A Figura 14 apresenta a média de sentimentos dos tópicos da amostra adicionados às categorias do Semantria.

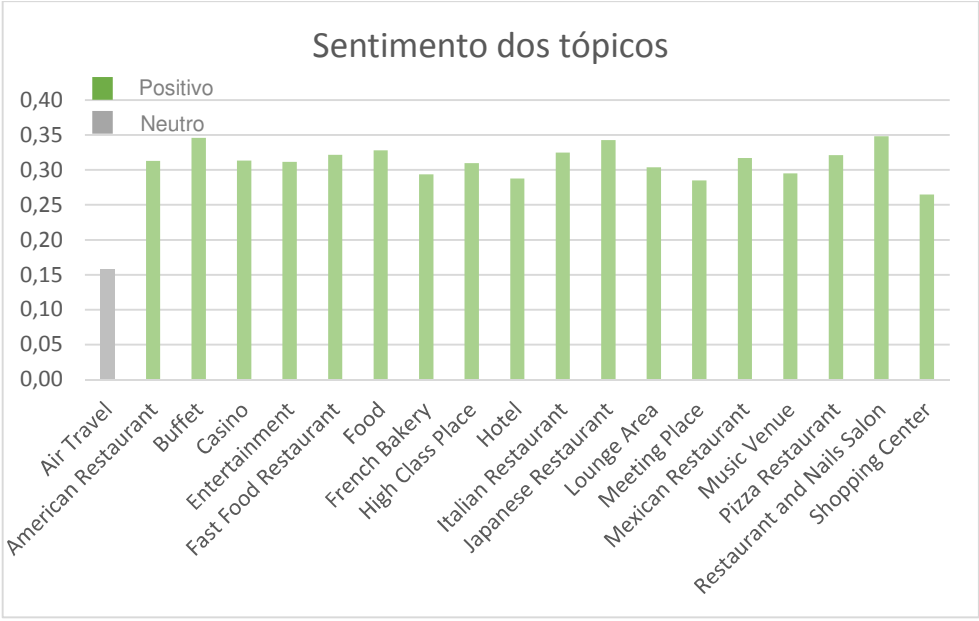

Figura 14 – Média de Sentimentos por Tópicos

Como se pode verificar na figura acima, em termos médios, a amostra em análise não apresenta tópicos com sentimentos negativos. Mas, uma vez que o intervalo para classificar os sentimentos positivos varia entre 0,5 e 10 o sentimentos dos tópicos não é assim tão positivo como parece sendo o tópico mais positivo o "Buffet" e o "Restaunt and Nails Salon" com 0,35 pontos. Os sentimentos neutros variam entre -0,5 e 0,22 e estão mais representados no tópico "Air Travel".

As figuras de 15 a 17 apresentam exemplos de entidades correlacionadas com os dois tópicos mais positivos em média e o tópico neutro contemplando as diferentes classificações de sentimentos.

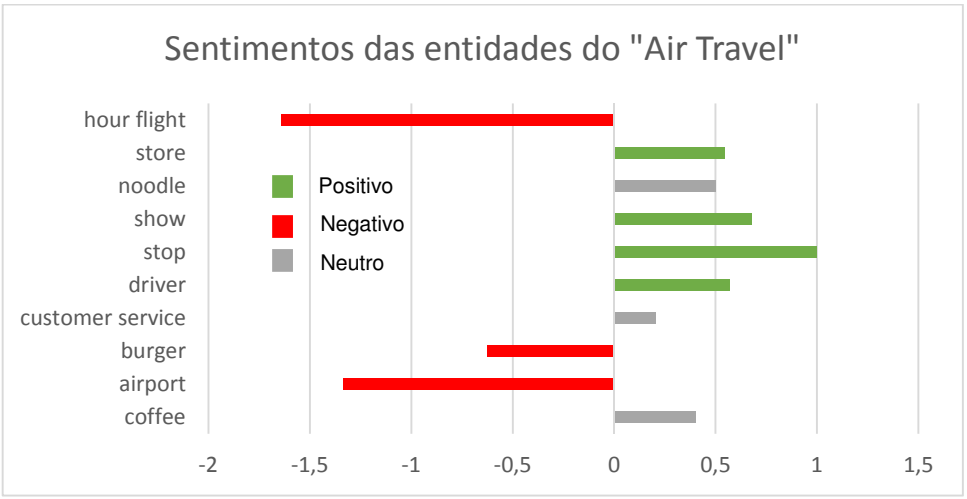

Figura 15 - Exemplo de sentimentos de algumas entidades do tópico "Air Travel"

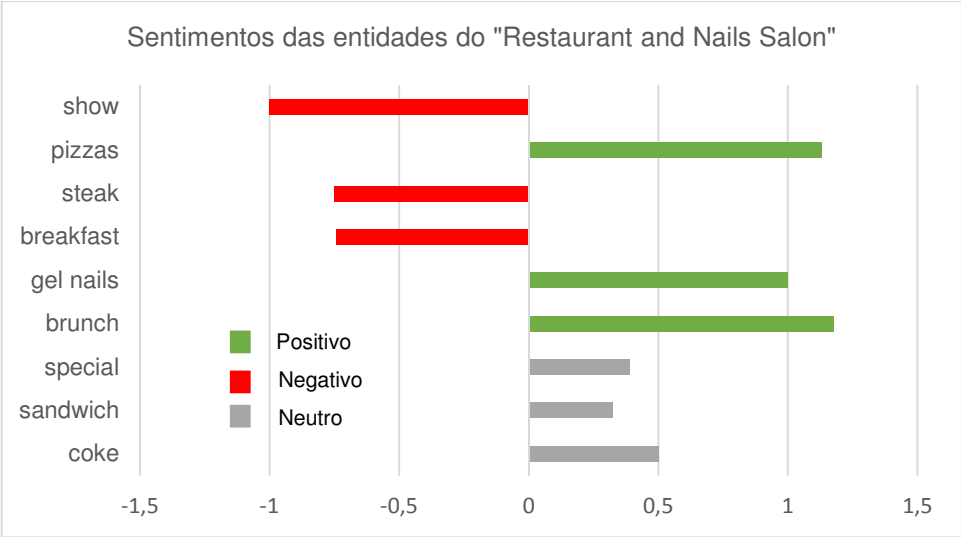

Figura 16 - Exemplo de sentimentos de algumas entidades do tópico "Restaurant and Nails Salon"

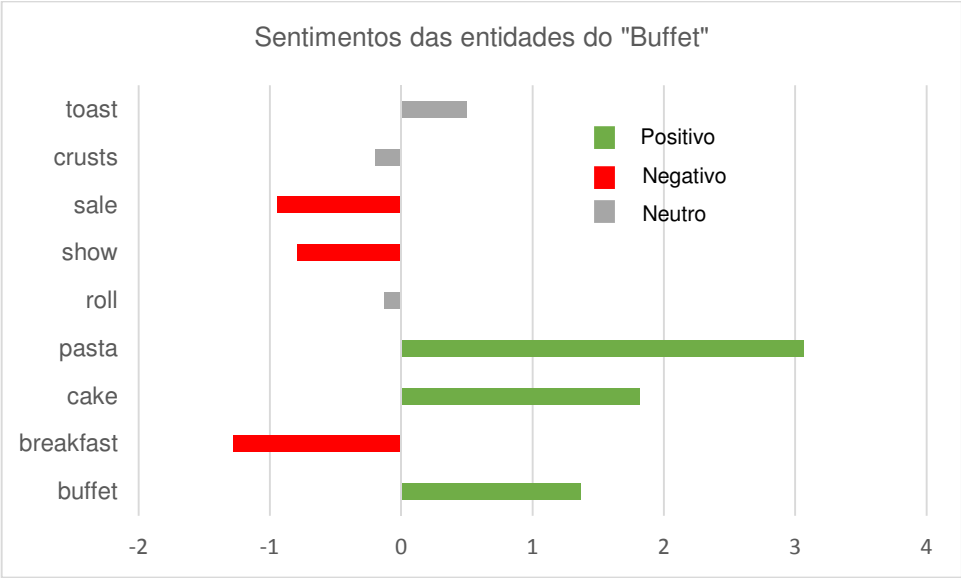

Figura 17 - Exemplo de sentimentos de algumas entidades do tópico "Buffet"

Terminada a análise de sentimentos, a DTM foi convertida em *data.matrix* para que seja possível adicionar os resultados obtidos, a fim de enriquecer o *dataset* para a construção do modelo. Assim, o dataset *input* final ficou constituído por 72 variáveis, como se pode observar na Tabela 5.

Tabela 5 – *Dataset input*

Dataset input	
Nome da variável	Descrição
sentiment	O *overall* do sentimento do comentário
votes_useful	Número total de votos *useful*
lenght	Número total de carateres que compõe o comentário

Dataset input	
Nome da variável	Descrição
topic	O número do tópico mais correlacionado com o comentário
entidades	A frequência dos 68 termos mais frequentes do DTM

Este *dataset* final ainda foi dividido em dois subconjuntos principais. Sendo um conjunto criado para treino com 70% dos dados e o outro conjunto criado para teste com os restantes 30% dos dados. Enquanto o conjunto de treino é necessário para construir o modelo, o conjunto de teste, como não tem influência no processo de aprendizagem, é o mais apropriado para testar os resultados obtidos pelos modelos.

3.4 Modelação

Depois de estruturar os dados em DTM e adicionar mais elementos para a análise (Tabela 5), estes foram utilizados para extrair padrões dos comentários que sejam passíveis de explicar a utilidade dos comentários (Feinerer *et.al.,* 2008). As principais tarefas de extração de conhecimento em *text mining* são a classificação, o *clustering*, a associação e a análise de tendências (Sharda *et.al.,* 2014, pp.248-252). Existem duas abordagens de classificação de texto que são a engenharia de conhecimento e aprendizagem automática (Feldman & Sanger, 2007 citado por Sharda *et.al.,* 2014, pp.248). Seguindo a abordagem da engenharia de conhecimento, o conhecimento de um especialista sobre as categorias é codificado declarativamente no sistema ou sob a forma de regras de classificação processuais. Através da abordagem por aprendizagem automática, um processo indutivo geral constrói um classificador, aprendendo com um conjunto de exemplos reclassificados. A tarefa a desenvolver do presente projeto é a classificação seguindo a abordagem de aprendizagem automática, por ser uma das técnicas mais utilizadas na descoberta de conhecimentos em *text mining* (Sharda *et.al.,* 2014, pp.248-252). Os modelos a criar são a Rede Bayesiana, o *Support Vector Machines (SVM), a* Regressão Logística e a Árvore de decisão.

3.4.1. Rede Bayesiana

A rede bayesiana usa o teorema de *Bayes*, que deve o seu nome a Thomas Bayes baseado em probabilidades condicionais (Larose, 2006, p.204-236):

$$p(Decisão_i\,|x) = \frac{p(x|Decisão_i)p(Decisão_i)}{p(x)} \qquad (1)$$

A regra de Bayes mostra como a variação das probabilidades hipótese (ou decisão) varia, tendo em conta novas evidências. Os requisitos necessários para a aplicação do teorema de Bayes como classificador requerem:

- Conhecer as probabilidades *à priori* $p(Decisão_i\ |x)$
- As probabilidades condicionais $p(x|Decisão_i)$

Este algoritmo assume que os atributos são independentes, mas tem demonstrado um bom desempenho mesmo em situações onde se encontram claras dependências entre atributos. Tem um comportamento bastante robusto ao ruído e a atributos irrelevantes. Toda a informação necessária à construção do modelo é adquirida com uma única passagem pelos dados, o que o torna um dos algoritmos de classificação mais rápidos.

As redes bayesianas são um tipo de modelo muito utilizado nas áreas onde são necessários realizar diagnósticos. Por exemplo, para um diagnóstico médico, um paciente informaria os seus sintomas, e com base nisso, um *software* pesquisaria na sua base de dados, alimentada com vários dados estatísticos pré-definidos ou adquiridos ao longo de seu uso, e retornaria possíveis doenças relativas aos sintomas apresentados, com certos níveis de probabilidade para cada doença. Um outro exemplo de utilização das redes bayesianas centra-se nos sistemas de *e-mail* onde muitas vezes é necessário aplicar filtros contra *spam,* e para isso é calculada a probabilidade de uma mensagem recebida ser *spam* ou não.

3.4.2. SVM

Apoiando-se na Teoria Estatística da Aprendizagem, as *Support Vector Machines* (SVM), surgiram pela primeira vez na década de setenta do século passado com o trabalho de Vapnik em 1979 (Vapnik, 2006), mas só recentemente a comunidade científica lhes tem dado mais importância. Tratam-se de algoritmos que podem ter algumas variações mediante a aplicação. Criados inicialmente para classificação de dados, recentemente começaram a ser aplicadas também em regressão (Vapnick *et.al.,* 1996). Têm um forte fundamento teórico e, também um bom desempenho na construção de modelos para conjunto de dados com muitos atributos e de registos.

As SVMs usam o *Kernel* para projetar os dados de entrada num espaço onde é esperado que a separação das classes seja mais fácil. Normalmente esse espaço apresenta maior dimensionalidade do que o espaço original. Outra característica importante é que o algoritmo das SVM tenta separar as classes através da maximização das margens e não pela minimização do erro. A Figura 18 apresenta um exemplo desta separação (Vapnick *et.al.,* 1996):

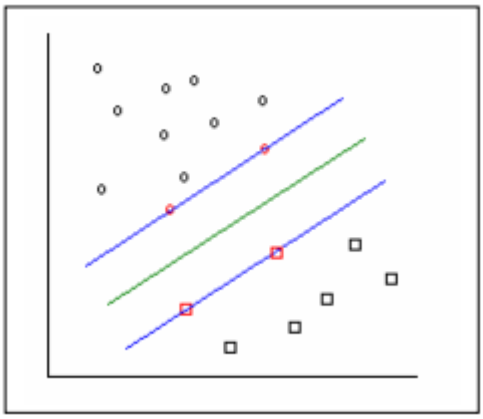

Figura 18 – Exemplo de separação das classes através da maximização das margens

Os dois *Kernels* mais comuns são o *Kernel* linear e *Kernel Gaussiano* (não-linear). Diferentes tipos de Kernels e diferentes escolhas dos seus parâmetros podem gerar diferentes propostas para os limites das margens. A utilização de *Kernels* não lineares permite obter fronteiras não lineares com um algoritmo que determina uma fronteira linear.

3.4.3. Regressão logística

Hosmer e Lemeshow (1989) apresentam o modelo de regressão logística, que também é conhecido por modelo logístico. Apresentam-no como um método de classificação utilizado no esclarecimento de problemas relacionados com a classificação dicotómica em várias áreas do conhecimento. Este modelo estabelece uma relação entre a probabilidade de ocorrência dos resultados de uma variável-resposta dicotómica, que normalmente é representada pelos termos “sucesso” e “fracasso”, “bom” e “mau” ou “useful” e “unusefull” e variáveis explicativas, sendo estas categóricas ou contínuas.

Dessa forma, considerando Y como a variável de classificação, sendo a categoria de sucesso igual a 1, e X_i como variável explicativa, a probabilidade de sucesso para a variável de classificação é dada pelo modelo indicado na seguinte equação (Larose, 2006, p. 155-199):

$$P(Y = 1) = \frac{\exp\{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\}}{1 - \exp\{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\}} \qquad (2)$$

Assim, através da probabilidade expressa por P(Y=1), é determinada a classificação ou não de um indivíduo como possuidor da característica em estudo. Geralmente, um ponto de corte é especificado para tal decisão. Um método frequentemente utilizado para estimar este ponto é a curva ROC.

A curva ROC, introduzida em 1993 por Zweig e Campbell, pode ser definida, geometricamente, como um gráfico em que para a abscissa tem-se a medida de 1-especificidade e para a ordenada tem-se a medida de sensibilidade, sendo esse plano designado unitário, pois cada eixo possui tamanho 1.

A sensibilidade é responsável pela proporção de indivíduos com a característica do modelo e a especificidade é responsável pela proporção de indivíduos sem a característica de interesse que é identificada corretamente pelo modelo. Assim, a curva ROC é construída variando o ponto de corte de classificação e através da amplitude das *scores*, para ambos os casos temos as *scores* como probabilidades.

3.4.4. Árvore de decisão

Uma técnica desenvolvida inicialmente pela Universidade de Michigan, com origem na área de aprendizagem automática, as Árvores de Decisão são representações gráficas das regras de classificação (Quinlan, 1993). A estrutura que cada representação segue é constituída por (Larose, 2005, p.107-126):

- Nó raiz – nó com o primeiro teste;
- Nós internos – cada um possui um teste a um atributo dos dados e têm duas ou mais sub-árvores que correspondem às respostas possíveis;
- Ramos – contendo valores dos atributos e
- Folhas – que representam as classes.

Os algoritmos de indução de árvores de decisão vão construindo a árvore de modo recursivo a partir de dados de treino, dividindo os dados em subconjuntos até que esses representem uma só classe ou respeitem determinados critérios.

A árvore de decisão tem a vantagem de ser construída de forma simples e rápida, é adequada para problemas com muitas dimensões e apresenta uma fácil representação e visualização. Para o construir são necessários muitos dados se o objetivo for descobrir estruturas complexas. Os principais algoritmos deste modelo são: CART, CHAID, AID, See5, ID4, ID6, C4.5 e C5 (Larose, 2005, p. 107). No presente projeto será utilizado o algoritmo C5.0 que é uma evolução do C4.5 apresentando pequenas diferenças como o método *boosting*.

Nas árvores de decisão é frequente surgir problemas de *overfitting,* que consiste no ajuste demasiado do modelo aos dados de treino. Quando o conjunto de treino não possui ruído, o número de erros no treino pode ser zero, mas quando este conjunto, entretanto, possui ruído, ou quando o conjunto de treino não é representativo, este algoritmo pode produzir árvores em que há *overfitting* (Larose, 2005, p.92).

Este problema muitas vezes é solucionado com a "poda" da árvore (Larose, 2005, p.115). A "poda" pode ser feita durante a aprendizagem, mas isto, muitas vezes, torna o processo mais complexo. Além disso, determinadas "podas" só podem ser decididas após a construção da árvore, pelo que a maior parte dos algoritmos faz a "poda" no final da construção da árvore. Após a "poda", surge o problema de determinar taxas de erro da árvore. Se houver dados em elevado número, pode-se reservar parte deles para teste após a construção da árvore (técnica de *reduced-error pruning*). Caso os dados sejam escassos, pode-se utilizar um esquema de *Cross-Validation* (Larose, 2005, p.105). Neste caso, os dados são divididos em N blocos de dimensão semelhante. A aprendizagem faz-se com

recurso a N iterações, em que a cada iteração são utilizados N-1 blocos para aprendizagem e a outra para teste, sendo este diferente a cada iteração.

3.4.5. Desenvolvimento

Para a fase de modelação recorreu-se à IBM SPSS Modeler 17.0[8] não só pelas suas múltiplas funcionalidades, mas também pela sua flexibilidade. Esta versão mais recente é muito acessível e "*user friendly*". Os modelos construídos foram parametrizados segundo a Tabela 6.

Tabela 6 – Parametrização dos modelos

Modelo	Parâmetros	Resultados	Variáveis mais importantes
Rede Bayesiana	✓ Structure type: **TAN** ✓ Method: **Maximum likelihood**		- steak - breakfast - appoint - stop - servic food
SVM	✓ Mode: **expert** ✓ Stopping criteria: **1.0E-3** ✓ Regularization parameter (C): **10** ✓ Regression precision (epsilon:) **0.1** ✓ Kernel type: **RBF** ✓ RBF gamma: **0.1**		- special - ice scream - breackfast - office - brunch
Regressão logística	✓ Procedure: **Binomial** ✓ Method: **Forwards**	✓ Cox and Snell = **0.135** ✓ Nagelkerke = **0.179** ✓ McFadden = **0.104**	- chair - valley - neighbordho o - place food
Árvore de decisão (C5.0)	✓ Output type: **Decision tree** ✓ Mode: **Simple** ✓ Use boosting **(10)** ✓ Cross-validate **(10)** ✓ Favor: **Accurary**	✓ Profundidade: **8**	- chair - burger - book - stop - appoint

[8] http://www-01.ibm.com/software/analytics/spss/products/modeler/

Como já se verificou na Figura 6, a variável *votes_useful* não é equilibrada e o mesmo problema também é verificado nas outras variáveis que compõe o *dataset.* Chawla (2009) defende que a distribuição natural, regra geral, não é a melhor distribuição para aplicar a um modelo de classificação. Além disso, o desequilíbrio nos dados pode ser mais característica da "escassez" (*sparseness*) no espaço característico do que o desequilíbrio das variáveis. Várias estratégias de reamostragem têm sido implementadas para solucionar este tipo de problemas; entre elas encontra-se o *random oversampling* e *random undersampling* (Chawla, 2009). No desenvolvimento do presente projeto será aplicada a estratégia do *random undersampling* com o intuito de melhorar os resultados do modelo.

No início da construção do modelo importou-se o *dataset* para o Modeler e de seguida, procedeu-se à transformação da variável *target.* Através do nó *Reclassify* dividiu-se a variável *votes_useful* em três níveis mediante os seguintes critérios:

- $0 \rightarrow 0$
- $[1, 3] \rightarrow 1$
- $[4, +\infty[\rightarrow 2$

A variável transformada foi designada de V3 e passou a ter a distribuição apresentada na Figura 19. O nível 0 e o nível 1 ficaram mais próximo um do outro com 42% e 50% registos associados, respetivamente. O nível 2, apesar de carregar os votos com valores mais elevados, contém os menos frequentes na amostra, o que fez com que o nível 2 ficasse mal representado, em relação aos dois primeiros níveis, com apenas 8% dos votos.

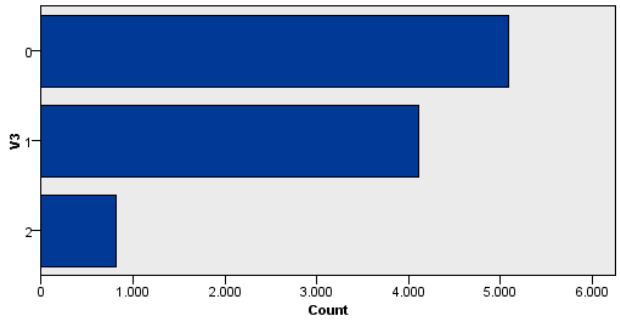

Figura 19 - Distribuição da variável target transformada (V3)

Ao aplicar os modelos apresentados na Tabela 7, verificou-se que a regressão logística e a árvore de decisão apresentavam taxas de erro muito elevados, quer na aprendizagem do modelo quer no teste, sendo que os piores resultados se verificavam na fase de testes. Mas verificou-se que não estão a fazer *overfitting*, e tanto o valor do *accuracy* como do *precision* em fase de treino e de teste estão muito perto um do outro. A rede bayesiana apresenta uma taxa de *accuracy* de aproximadamente 78% na fase de treino; no entanto, quando é aplicado sobre os dados de teste, a taxa do *accuracy* baixa para 46%, o que significa que o modelo não consegue prever corretamente para novos dados. O mesmo se verificou com o modelo SVM que apresentou um ótimo resultado na aprendizagem do modelo, com uma taxa accuracy de aproximadamente 98% mas os valores caíram para 42% quando

o modelo foi executado com os dados de teste. Pensa-se que isto se deve ao facto de estes dois últimos modelos estarem a fazer *overfitting*.

Tabela 7 - Medidas de desempenho - *target* com 3 níveis

Modelos	Dados	Accuracy (%)	Precision (%)	Erro médio (%)
Logistic	Training	59,46	60,62	40,54
	Testing	48,94	53,19	51,06
Bayes Net	Training	**77,61**	**78,58**	22,39
	Testing	46,02	**52,44**	53,98
C5.0	Training	55,59	55,44	44,41
	Testing	51,71	52,73	48,29
SVM	Training	**97,66**	**97,25**	2,34
	Testing	42,18	**53,86**	57,82

Após algumas experiências verificou-se que a variável *Length* não estava a contribuir para melhorar o modelo e foi excluída da análise. Ainda no sentido de melhorar os resultados e tentar corrigir algum do desfasamento da variável *target* em 3 níveis, fez-se o balanceamento dos dados. A partir do nó *Distribuition* criou-se o gráfico de distribuição da variável V3 e depois foi gerado o nó *Balance Node (reduce),* levando a que a variável passasse a ter a distribuição apresentada na Figura 20.

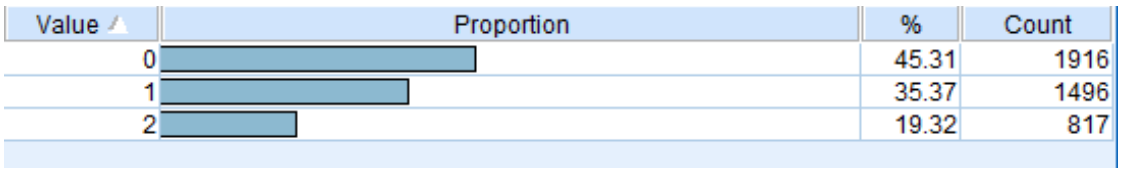

Figura 20 - Distribuição da variável target balanceada

O balanceamento foi feito sobre os dados de treino (Yap *et.al.,* 2014), o nó *Balanced* foi ligado ao nó *Partition,* e de seguida foi exportado para um *flat file* onde foram conectados os nós dos modelos. Nesta análise os modelos apresentaram uma taxa de *accurracy* não inferior a 73% na fase de aprendizagem, mas na fase dos testes as taxas ficaram abaixo dos 34%, como se pode verificar na Tabela 8.

Tabela 8 – Medidas de desempenho dos modelos aplicados sobre os dados balanceados

Modelo	**Dados**	**Accuracy (%)**	**Precision (%)**	**Erro médio (%)**
Logistic	Training	73,46	69,73	26,54
	Testing	30,61	52,04	69,39
Bayes Net	Training	92,04	91,08	7,96
	Testing	33,19	51,50	66,81
C5.0	Training	**74,13**	**70,62**	25,87
	Testing	33,55	**54,90**	66,45
SVM	Training	**99,27**	**99,46**	0,73
	Testing	27,65	**52,48**	72,35

Ainda na tentativa de conseguir um modelo com melhores resultados do que já foi apresentado, decidiu-se fazer uma nova reclassificação da variável *votes_useful* (Figura 6) considerando que todos os votos a zero continuavam a ser 0 e que qualquer valor acima seria considerado 1, criando assim uma variável dicotómica (Hosmer & Lemeshow, 1989) designada "useful", com uma distribuição mais equilibrada como apresenta a Figura 21.

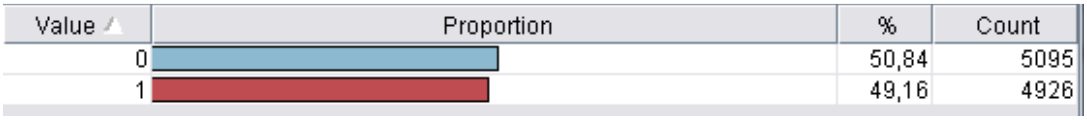

Figura 21 – Distribuição da variável *target* transformada numa variável binária

Com este processo notou-se uma ligeira diferença nos números, mas o mesmo comportamento nos modelos, ou seja, os mesmos pares de modelos (SVM e Bayes Net) apresentaram uma alta taxa de *accurracy* na fase de treino, mas baixa taxa de *accuracy* na fase de teste, enquanto os outros dois (Logistic e C5.0) continuaram a apresentar baixo *accuracy* (menos de 60%); no entanto, a diferença entre as duas fases (aprendizagem e teste) é mínima. A Tabela 9 mostra a comparação das taxas de *accuracy* apresentados pelos modelos, quer na fase de teste quer na fase de treino para as três transformações efetuadas ao longo do desenvolvimento do projeto.

Tabela 9 - Comparação das taxas *accuracy* entre os modelos e as variáveis *target* transformadas

Modelo	Fase	Accuracy (V3)	Accuracy (V3 balanceado)	Accuracy (useful)
Logistc	Treino	59,46	73,46	63,10
	Teste	48,94	30,61	53,90
Bayes net	Treino	77,61	92,04	79,17
	Teste	46,02	33,19	52,10
C5.0	Treino	55,59	74,13	56,84
	Teste	51,71	33,55	53,41
SVM	Treino	97,66	99,27	97,21
	Teste	42,18	27,65	52,55

3.5 Avaliação

Após a extração dos padrões, os modelos devem ser avaliados e interpretados no domínio do contexto do problema, a fim de verificar se os padrões produzidos são válidos e úteis ao objetivo final do processo (Chawla, 2009). Na validação de tarefas preditivas, pode-se, por exemplo, verificar a precisão do modelo através dos valores indicados no conjunto de dados de teste (Chawla, 2009).

Um modelo de classificação, geralmente, é avaliado através da matriz de confusão (Figura 22). Esta quantifica quantos exemplos do *dataset* seriam bem classificados pelo modelo construído (representado na diagonal principal) sendo que os outros seriam os mal classificados. A coluna representa a situação prevista e a linha representa a situação atual (Chawla, 2009). É sobre esta matriz que se calculam as medidas necessárias para avaliar o desempenho dos modelos.

		Previsto		Total
		Useful (+)	Unuseful (-)	
Atual	Useful (+)	Verdadeiro Positivo (hit) - **VP**	Falso Negativo (miss, erro tipo II) - **FN**	VP + FN = P
	Unuseful (-)	Falso Positivo (falso alarme, erro tip I) - **FP**	Verdadeiro Negativo (rejeição correta) - **VN**	FP + VN = N
Total		VP + FP = N_1	FN + VN = P_1	

Figura 22 - Matriz de confusão

Fonte: Adaptado de Chawla, 2009; Yap, Rani, Rahman, Fong, Khairudin, & Abdullah, 2014; Liu, 2008

O método mais utilizado para avaliar um modelo é a precisão (*accuracy*), mas vários autores defendem que, para *datasets* desequilibrados (como o deste projeto), esta pode não ser a medida mais apropriada (Chawla, 2009; Yap *et.al.,* 2014; Liu, 2008). Por isso, para além do *accuracy*, a *sensitivity*, a *precision*, a *specificity* e o *F-score* (Yap et. Al. 2014) foram os escolhidos como critério para medir o desempenho dos modelos. O principal objetivo de aprender a partir de conjuntos de dados desequilibrados é melhorar a *sensitivity* sem prejudicar a *precision* (Chawla, 2009). No entanto, os objetivos de *sensitivity* e *precision* podem ser muitas vezes conflituantes, uma vez que quando se aumenta o verdadeiro positivo para o caso minoritário, o número de falsos positivos também pode ser aumentado e isto reduzirá a *precision* (Chawla, 2009). A métrica *F-score* é uma medida que combina os *trade-offs* de *precision* e *sensitivity* e emite um único número que reflete a "bondade" de um modelo na presença de classes raras (Chawla, 2009). Os resultados da avaliação do modelo são apresentados na Tabela 10.

3.5.1. Medidas de desempenho

- *Accuracy* – Percentagem da amostra que foram corretamente classificados pelo modelo.

$$a = \frac{VP + VN}{P + N} \quad (3)$$

- *Specificity* – Mede o quão bom o modelo é a acertar nos negativos

$$sp = \frac{VN}{N} \quad (4)$$

- *Recall/Sensitivity* – Mede o quão bom o modelo é a acertar nos positivos

$$r = \frac{VP}{P} \quad (5)$$

- *Precision* – Mede o quão bom o modelo é a acertar nos positivos classificados como positivos

$$p = \frac{VP}{N_1} \quad (6)$$

- F-*score* – A média "harmónica" entre a *precision* e a *sensitivity*

$$r = \frac{2pr}{p + r} \quad (7)$$

- Taxa de erro – Mede o erro tradicional do modelo

$$e = \frac{FP+FN}{P+N} \quad (8)$$

Orientando-se então pelos resultados do cálculo das medidas apresentadas na Tabela 10 nota-se que nenhum dos modelos aplicados apresenta pouca coerência entre a aprendizagem e o teste. O SVM e o *Bayes Net* continuam a ser os modelos com melhores valores representativos com taxas de *sentivity* e F-score acima dos 79%, uma taxa aceitável na escolha de um bom modelo. Porém, estes apenas dizem respeito à fase de aprendizagem e em relação aos testes estes caem para taxa entre 53% e 54%. Por seu turno, a árvore C5.0 apresenta valores muito baixos de *sensitivity* quer na fase de aprendizagem como na fase de teste (valores abaixo dos 40%) e apresenta um F-score que não chega sequer aos 50 % (é o que apresenta piores resultados). Por fim, o *logistic* apresenta 67% de *sensitivity* e 65% de *F-score* na fase de aprendizagem e na fase de teste apresenta 58% de *sensitivity* e 57% de *F-score*.

Tabela 10 – Medidas de desempenho dos modelos aplicados sobre o *dataset* com a *target* "useful"

Modelo	Dados	Accuracy (%)	Sensitivity (%)	Especificity (%)	Precision (%)	F-score (%)	Erro médio (%)
Logistic	Training	63,10	**67,01**	59,16	62,28	**64,56**	36,90
	Testing	53,90	**58,15**	49,18	55,93	**57,02**	46,10
Bayes Net	Training	79,17	80,39	77,94	78,58	79,48	20,83
	Testing	52,10	52,59	51,55	54,63	53,59	47,90
C5.0	Training	56,84	39,53	74,27	60,72	47,88	43,16
	Testing	53,41	36,20	72,37	59,08	44,89	46,59
SVM	Training	97,21	97,46	96,96	96,99	97,22	2,79
	Testing	52,55	46,51	59,21	55,68	50,68	47,45

Os resultados não chegam a atingir os valores médios considerados para um bom modelo, mas pelos valores do F-score, se fosse para escolher o melhor modelo seria o *logistic*, por prever os dados

de teste com a mesma precisão que apreendeu o modelo. A Figura 23 apresenta as variáveis que foram mais importantes na construção do modelo regressão logística.

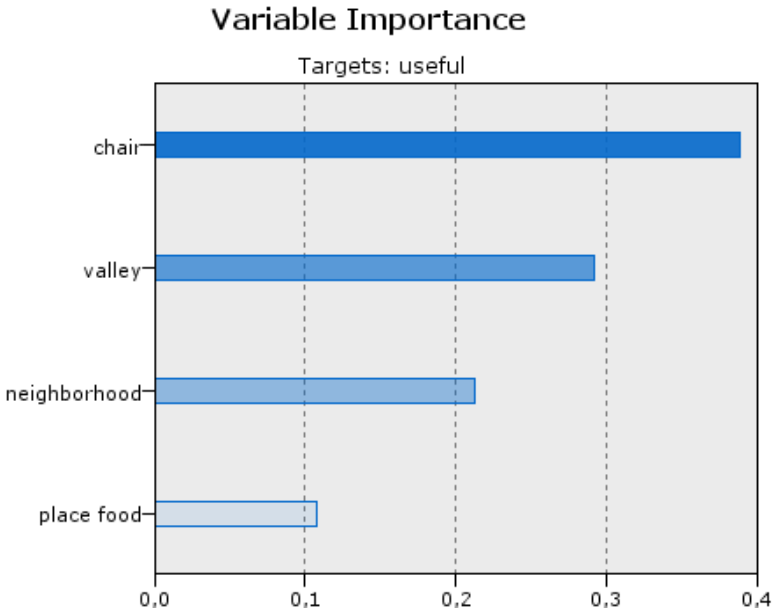

Figura 23 – Importância das variáveis no resultado da regressão logística

4 Análise dos Resultados

A análise dos tópicos revelou que o tópico "Buffet" é, em termos médios e percentuais, o mais positivo e mais útil de entre os tópicos, o que permite dizer que, os termos que estão mais correlacionados com este tópico são os que devem ser considerados na classificação de um comentário útil. E uma vez que os tópicos construídos são pouco homogéneos e porque em todos os tópicos se encontram todos os termos em análise, apresentados por ordem de correlação, sugere-se que os termos mais correlacionados com o tópico "Buffet" e que aparecem como sendo os mais correlacionados com outros tópicos, permitem classificar estes tópicos como sendo úteis também.

O top 5 dos termos mais correlacionados com o tópico 1 são: **Buffet, Breakfast, Cake, Roll** e **Pasta**. O termo **Buffet** figura como um elemento do top 5 do tópico 15 "Meeting Place" que ainda apresenta os termos Store, Burger, Mail e Game. O termo **Breakfast** figura no top 5 dos seguintes tópicos:

i. Tópico 2 - "American Restaurant" juntamente com os termos *Burger, coffee, Ice cream* e *Sandwich* – este tópico foi o segundo com maior número de votos "useful";
ii. Tópico 9 - "Restaurant and Nails Salon" em conjunto com os termos *Sandwich, Pizza, Nail, Breakfast* e *show* – este tópico apresentou o segundo valor médio mais alto de sentimento positivo.

O termo **Roll** faz parte do tópico 8 "Japanese Restaurant" juntamente com *Sushi, Sushi place, Sandwich* e *Sushi bar* por sua vez os termos **Pasta** e **Cake**, não aparecem em mais nenhum top 5 dos tópicos construídos.

Olhando para os resultados da Regressão Logística, nota-se que os termos mais importantes para classificar um comentário como sendo útil são: *chair, valley, neighborhood* e *place food*. Sugere-se assim, que os comentários relacionados com lugares de refeição (*place food, chair*) como restaurantes, bares, pastelarias, etc. perto de Valley, têm uma maior probabilidade de serem considerados úteis.

Uma vez que alguns comentários fazem referência ao Valley relacionando-o com Las Vegas, assumiu-se que Valley é o mesmo que Las Vegas Valley, e sendo a cidade de Las Vegas um dos primeiros destinos turísticos do mundo, muito movimentada e cheio de atrações (Okada, 2014), é normal que os comentários que falam dela sejam consideradas úteis.

5 Conclusão

Este estudo apresenta a importância da análise de sentimentos e do *text mining* em geral na extração de conhecimentos de dados não estruturados e suporte à tomada de decisão. Recorreu-se ao *software* R para a fase de pré-processamento onde foi realizada a limpeza e transformações, e estruturação dos dados, bem como algumas análises descritivas básicas como a frequência dos termos, que resultou na construção do *wordcloud* e a correlação entre os termos, um primeiro passo para a construção dos tópicos. Ainda no R executou-se o *part-of-speech* de onde foram extraídas as entidades que em conjunto com os tópicos e os comentários, foram processados no *software* Semantria para a extração de sentimentos escondidos em cada um deles.

A estruturação dos dados resultou na matriz DTM com a frequência relativa dos termos sobre os documentos. Devido à grande dimensão da DTM não foi possível extrair todos os dados do R, pelo que o ficheiro .csv exportado para ser utilizado no Modeler apenas continha os termos mais frequentes, num total de 68 termos. Novas variáveis como os tópicos e sentimento (*overall*), extensão do comentário (*length*) e votos "useful" foram adicionados ao DTM, constituindo-se assim o *dataset input* para a construção do modelo.

Apesar da análise efetuada, foram identificadas algumas limitações e potenciais investigações futuras. Os tópicos construídos apresentaram pouca consistência e olhando apenas para os termos mais correlacionados com cada um dos tópicos, nota-se que os tópicos não são completamente heterogéneos. Devido a isso, foi necessário um esforço redobrado para investigar o conteúdo de alguns comentários relacionados com cada um dos tópicos e fazer uma análise comparativa dos tópicos com as categorias de negócio da Yelp para ajudar no processo de *profiling.* Mesmo assim não foi possível dar um nome distinto e claro a todos os tópicos devido à mistura de termos e assuntos.

Na análise de sentimentos o Semantria devolveu o sentimento (*overall*) de cada comentário e de cada entidade e tópicos relacionados com cada um dos comentários. Esta particularidade é muito importante, dado que muitas vezes se encontram comentários que fazem referência às várias entidades e tópicos onde o sentimento do comentário, no geral, pode ser positivo, mas as entidades ou os tópicos não o são. E é muito importante haver esta separação na análise. Um outro elemento de análise que o Semantria também disponibiliza é o portador de sentimento: a chave que faz com que o sentimento seja positivo ou negativo. O Semantria é uma ferramenta de grande utilidade na área de análise de sentimento com uma funcionalidade interessante e de fácil entendimento. Mas como todo o *software*, este também não é perfeito e detetaram-se algumas falhas a nível do *part-of-speech* e da classificação das negações, problemas comuns em vários algoritmos que, com o tempo, vão sendo solucionados.

A construção do modelo revelou uma qualidade indesejada dos dados da amostra, sendo uma das causas apontadas pela existência de uma taxa de erro médio muito elevada dos modelos, onde nenhum dos modelos construídos apresentaram uma taxa de *accuracy* acima dos 70% em ambas as fases (treino e teste). Uns apresentavam baixas taxas de *accuracy* tanto em dados de treino como de teste; outros faziam *overfitting* apresentando uma boa taxa de *accuracy* para os dados de treino, mas

uma taxa muito baixa para os dados de teste. Foram ainda calculadas as outras medidas de avaliação de desempenho como a *precision* e a *sensitivity,* mas continuou a ser difícil melhorar a precisão do modelo. No entanto, a escolha final recaiu sobre a regressão logística por ser a que apresenta menos desfasamento entre os dados de treino e de teste.

Os comentários publicados nas plataformas de recomendações *online* como a Yelp são classificados como útil pelos próprios consumidores, que vão à procura de informação, com um voto "useful". É com base neste voto que o sistema de recomendação da Yelp consegue identificar quais os comentários mais úteis para os consumidores e posteriormente colocá-los em destaque, a modo a ajudar e influenciar mais consumidores na sua tomada de decisão, e ajudar também as empresas a aumentar as suas vendas, a angariar mais clientes e a ganhar conhecimentos para as suas tomadas de decisão de marketing. Com isto, é respondida a questão de Litvin, Goldsmith e Pan (2008) "Na ausência do contacto face-a-face com quem publica os comentários, quais os critérios utilizados pelo consumidor para determinar a confiabilidade das influências dos *social media*?". Não é necessário conhecer quem publicou o comentário para saber se ele é confiável e útil, pois os próprios membros da comunidade é que determinam essa utilidade.

Tendo os resultados da análise de sentimentos e do modelo, a Yelp consegue melhorar o seu sistema de recomendação e colocar em destaque os comentários que apresentam os termos que determinam a utilidade dos comentários, mesmo antes do novo comentário receber qualquer voto.

Uma vez que o âmbito do projeto passa pelo estudo dos comentários *online* ligados ao produto turístico, foram considerados todos os tipos de produtos e de negócio ligados ao turismo. Isto pode ter contribuído para que os resultados não fossem tão consistentes. Por isso, para o futuro recomenda-se a limitação de âmbito de pesquisa a, por exemplo, restaurantes, hotéis ou viagens.

6 Bibliografia

Abeywardena, I. S. (2014). Public opinion on OER and MOCC: A sentiment analysis of twitter data. Paper presented at the international conference on open and flexible education (ICOFE 2014). Hong Kong, China.

Abrahams, A. S., Jiao, J., Wang, G. A., & Fan, W. (2012). Vehicle defect discovery from social media. *Decision Support Systems, 54*(1), 87–97. Acedido dezembro 5, 2015, em http://doi.org/10.1016/j.dss.2012.04.005.

Adomavicius, G., & Tuzhilin, A. (2005). Toward the Next Generation of Recommender Systems: A survey of the State-of-the-art and Possible Extensions. IEEE Transactions on Knowledge and Data Engineering, 17, 734–749.

Alexa. (2015a). How popular is consumerreports.org? Acedido julho 25, 2015, em http://www.alexa.com/siteinfo/consumerreports.org.

Alexa. (2015b). How popular is tripadvisor.com? Acedido julho 25, 2015, em http://www.alexa.com/siteinfo/tripadvisor.com.

Alexa. (2015c). How popular is trivago.com? Acedido julho 25, 2015, em http://www.alexa.com/siteinfo/trivago.com.

Alexa. (2015d). How popular is yelp.com? Acedido julho 25, 2015, em http://www.alexa.com/siteinfo/yelp.com.

Asur, S., & Huberman, B. A. (2013). Predicting the Future with Social Media. *Applied Energy*, *112*, 1536–1543. doi:10.1016/j.apenergy.2013.03.027.

Bakhshi, S., Kanuparthy, P., & Shamma, D. A. (2015). Understanding Online Reviews: Funny, Cool or Useful? (pp. 1270-1276). ACM Press. Acedido junho 11, 2015, em http://doi.org/10.1145/2675133.2675275.

Berry, L. L. (2002). Relationship Marketing of Services Perspectives from 1983 and 2000. *Journal of Relationship Marketing, 1*(1), 59–77. Acedido junho 15, 2015, em http://doi.org/10.1300/J366v01n01_05.

Berthon, P. R., Pitt, L. F., Plangger, k. & Shapiro, D. (2012). Marketing meets Web 2.0, social media and creative consumers: Implications for international marketing strategy. *Business horizons, 55*(3), 261-271.

Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G., & Reynar, J. (2008). Building a sentiment summarizer for local service reviews. Acedido junho 10, 2015, em http://www.ryanmcd.com/papers/local_service_summ.pdf.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics, 1*(1), 17–35.

Blei, D. M, & Lafferty, J. D. (2009). Topic models. *Text mining: classification, clustering, and application, 10*(71), 34.

Bollen, J., Pepe, A., & Mao, H. (2009). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *arXiv:0911.1583 [cs].* Acedido fevereiro 20, 2015, em http://arxiv.org/abs/0911.1583

Borràs, J., Moreno, A., & Valls, A. (2014). Intelligent tourism recommender systems: A survey. *Expert Systems with Applications*, *41*(16), 7370–7389. doi:10.1016/j.eswa.2014.06.007.

Cavazza, F. (2011). Description des different types de medias sociaux. Acedido Julho 29, 2015, em http://www.mediassociaux.fr/2011/02/06/description-des-differents-types-de-medias-sociaux/.

Cavazza, F. (2015). Panorama des medias sociaux 2015. Acedido Julho 29, 2015, em http://www.fredcavazza.net/2015/05/29/panorama-des-medias-sociaux-2015/.

Chan, N. L., & Guillet, B. D. (2011). Investigation of Social Media Marketing: How Does the Hotel Industry in Hong Kong Perform in Marketing on Social Media Websites? *Journal of Travel & Tourism Marketing*, 28(4), 345–368. doi:10.1080/10548408.2011.571571.

Chawla, N. V. (2009). Data Mining for Imbalanced Datasets: An Overview. Em O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 875-886). Springer US. Acedido outubro 1, 2015, em http://link.springer.com/chapter/10.1007/978-0-387-09823-4_45.

Chopra, S., & Bangalore, S. (2012). Weakly supervised neural networks for Part-Of-Speech tagging (pp. 1965–1968). IEEE. Acedido maio 18, 2015, em http://doi.org/10.1109/ICASSP.2012.6288291.

Consumer Reports. (2015). About us. Acedido agosto 10, 2015, em http://www.consumerreports.org/cro/about-us/index.htm.

CRISP-DM. (2000). CRISP-DM 1.0: Step-by-step data mining guide. Acedido junho 11, 2015, em http://www.the-modeling-agency.com/crisp-dm.pdf.

Darwiche, A. (2010). Bayesian networks. Communications of the ACM, 53(12), 80. Acedido setembro 18, 2015, em http://doi.org/10.1145/1859204.1859227.

Deepthi.V, G., Rekha, K. S. (2014). Opinion mining and classification of User reviews in social media. *International Journal of advance Research in computer Science and management Studies, 2(4).* Acedido julho 25, 2015, em www.ijarcsms.com/docs/paper/volume2/issue4/V2I4-0005.pdf.

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science, 49*(10), 1407–1424.

Dwivedi, M., Shibu, T. P., & Venkatesh, U. (2007). Social software practices on the Internet: Implications for the hotel industry. *International Journal of Contemporary Hospitality Management, 19*(5), 415-426.

Feinerer, I., Hornik, K., Meyer, D. (2008). Text Mining Infrastruture in R. *Journal of Statistical Software. 25* (5). Acedido outubro 10, 2015, em http://www.jstatsoft.org.

Feldman, R., Fresko, M., Hirsh, H., Aumann, Y., Liphstat, O., Schler, Y., & Rajman, M. (1998). Knowledge Management: A Text Mining Approach. *Proceeding of the 2nd Internacional Conference on Pratical Aspects of Knowledge Management (PAKM98)* (pp. 9.1. – 9.10).

Feldman, S. (1999). NLP meets the jabberwocky: Natural language processing in information retrieval: Search Engine Section. (Online, Ed.). Information Today, Inc. Acedido junho 9, 2015, em http://www.scism.lsbu.ac.uk/inmandw/ir/jaberwocky.htm

Ferrão, F. (2003). *CRM - Marketing e Tecnologia*. Lisboa, Portugal: Escolar Editora.

Fotis, J. (2012). Discussion of the impacts of social media in leisure tourism: "The impact of social media on consumer behavior: Focus on leisure travel". Acedido maio 5, 2015 em http://johnfotis.blogspot.com.au/2012_03_01_archive.html.

Gillingan, C., Wilson, R. M. S. (2003). *Strategic Marketing Planning*. Butterworth-Heinemann, Oxford.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yoogatama, D., Fanigan, J., Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. Em *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2* (pp. 42–47). Association for Computational Linguistics. Acedido junho 3, 2015, em http://dl.acm.org/citation.cfm?id=2002747.

Godbole, S., Bhattacharya, I., Gupta, A., & Verma, A. (2010). Building Re-usable Dictionary Repositories for Real-world Text Mining. *Em Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (pp. 1189–1198). New York, NY, USA: ACM. Acedido fevereiro 16, 2015, em http://doi.org/10.1145/1871437.1871588.

Godes, D., & Mayzlin, D. (2004). Using Online Conversations to Study Word-of-Mouth Communication. *Marketing Science, 23*(4), 545–560. Acedido junho 11, 2015, em http://doi.org/10.1287/mksc.1040.0071.

Gopal, R., Marsden, J. R., & Vanthienen, J. (2011). Information mining — Reflections on recent advancements and the road ahead in data, text, and media mining. *Decision Support Systems, 51*(4), 727–731. Acedido junho 11, 2015, em http://doi.org/10.1016/j.dss.2011.01.008.

Graham, W. (2014). Hands-On Data Science with R Text Mining. Acedido outubro 14, 2015, em http://togaware.com/onepager/.

Grant-Braham, B. (2007). The social media and travel chatter. *Hospitality in Focus.* Acedido junho 10, 2015, em http://eprints.bournemouth.ac.uk/13634/2/Croner_-_Hospitality_in_Focus_-_14th_September_2007.pdf.

Greenberg, P. (2010). The impact of CRM 2.0 on customer insight. *The Journal of Business and Industrial Marketing, 25* (6), 410–419.

Griffiths, T. L. & Steyvers, M. (2004). Finding scientific topics. *In Proceedings of the National Academy of Sciences of the United States of America* (pp. 5228–5235).

Grün, B., Hornik, K. (2011). topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software 40*(13). Acedido julho 25, 2015, em http://www.jstatsoft.org/article/view/v040i13.

Guernsey, L. (2000). Suddenly, Everybody's an Expert. *The New York Times*. Acedido março 13, 2014, em http://www.nytimes.com/2000/02/03/technology/suddenly-everybody-s-an-expert.html.

Guerreiro, J., Rita, P., & Trigueiros, D. (2015). A Text Mining-Based Review of Cause-Related Marketing Literature. *Journal of Business Ethics*, 1–18. http://doi.org/10.1007/s10551-015-2622-4

Hajas, P., Gutierrez, L., & Krishnamoorthy, M. S. (2014). Analysis of Yelp Reviews. Acedido julho 25, 2015, em https://vpn2.iscte.pt/+CSCO+0h756767633A2F2F6E656B76692E626574++/abs/1407.1443.

Hearst, M.A. (1999). Untangling text data mining. *Proceeding of the 37th annual meeting of the Association for computational Linguistics on Computational Linguistics* (pp.3-10). Stroudsburg, PA, USA: Association for Computational Linguistics. doi: 10.3115/1034678.1034679.

Hochreiter, R., & Waldhauser, C. (2014). Data Mining Cultural Aspects of Social Media Marketing. *arXiv:1401.5726 [physics].* Acedido fevereiro 20, 2015, em http://arxiv.org/abs/1401.5726.

Hosmer, D. W. J., Lemeshow, S. (1989). *Applied logistic regression*. John Wiley, NY.

Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems, 52*(3), 674–684. Acedido junho 11, 2015, em http://doi.org/10.1016/j.dss.2011.11.002.

Huang, L. (2012). Social Media as a New Play in a Marketing Channel Strategy: Evidence from Taiwan Travel Agencies' Blogs. *Asia Pacific Journal of Tourism Research,* 17(6), 615–634. doi:10.1080/10941665.2011.635664.

Inversini, A., Cantoni, L., & Buhalis, D. (2009). Destinations' Information Competition and Web Reputation. *Information Technology & Tourism*, 11(3), 221–234. doi: 10.3727/109830509X12596187863991.

IRTS. (2008). International Recommendations for Tourism Statistics. Acedido abril 18, 2015, em http://unstats.un.org/unsd/publication/Seriesm/SeriesM_83rev1e.pdf#page=21.

Jain, M., Kumar, M., & Aggarwal, N. (2013). Web Usage Mining: An Analysis. *Journal of Emerging Technologies in Web Intelligence*, *5*(3), 240–246. doi:10.4304/jetwi.5.3.240-246.

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, *60*(11), 2169–2188. doi:10.1002/asi.21149.

Kaplan, A. M, Haenlein, M. (2010). Users of the world, unite! The challenges and Opportunities of Social Media. *Business Horizons 53*, 59-68.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. Business Horizons, 54(3), 241–251. doi:10.1016/j.bushor.2011.01.005.

Kozinets, R. V., Hemetsberger, A., & Schau, H. J. (2008). The Wisdom of Consumer Crowds: Collective Innovation in the Age of Networked Marketing. *Journal of Macromarketing, 28*(4), 339–354. Acedido março 3, 2015, em http://doi.org/10.1177/0276146708325382.

Larose, D.T. (2005). *Discovering knowledge in data: an introduction to data mining*. John Wiley, NY.

Larose, D.T. (2006). *Data mining: methods and models*. John Wiley, NY.

Lawrence, L. (2014). Reliability of Sentiment Mining Tools: A comparison of Semantria and Social Mention. Acedido outubro 13, 2015, em http://essay.utwente.nl/65302/.

Leung, D., Law, R., van Hoof, H., & Buhalis, D. (2013). Social Media in Tourism and Hospitality: A Literature Review. *Journal of Travel & Tourism Marketing*, *30*(1-2), 3–22. doi:10.1080/10548408.2013.750919.

Lindon, D., Lendrevie, J., Lévy, J., Dionísio, P., Rodrigues, J. V. (2009). *Mercator XXI.* (12th ed.). Lisboa, Portugal: Dom Quixote.

Linoff, G. S. & Berry, M. J. A. (2001). *Mining the web: transforming customer data into customer value*. John Wiley Sons, NY.

Litvin, S. W., Goldsmith, R.E., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. *Tourism Management, 29*(3), 458-468. Acedido março 3, 2015, em http://doi.org/10.1016/j.tourman.2007.05.011

Liu, B. (2008). *Web Data Mining: Exploring Hyperlinks,Contents, and Usage Data*. (2nd ed.). Springer Berlin Heidelberg New York.

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypoool Publishers.

Miller, T.W. (2005). *Data and text mining: a business applications approach*. Upper Saddle River, NJ: Pearson Education International.

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications, 40*, 4241-4251.

Munar, A. M. (2011). Tourist-created content: rethinking destination branding. *International Journal of Culture, Tourism and Hospitality Research*, *5*(3), 291–305. doi:10.1108/17506181111156989.

Munar, A. M., & Jacobsen, J. K. S. (2014). Motivations for sharing tourism experiences through social media. *Tourism Management, 43*, 46–54. Acedido fevereiro 19, 2015, em http://doi.org/10.1016/j.tourman.2014.01.012

Murphy, L., Moscardo, G., & Benckendorff, P. (2007). Using Brand Personality to Differentiate Regional Tourism Destinations. Journal of Travel Research, 46(1), 5–14. doi:10.1177/0047287507302371.

Okada, S. (2014). HVS - In Focus: Las Vegas Casino & Hotel Market Outlook 2014. Acedido outubro 20, 2015, em http://www.hvs.com/article/6968/in-focus-las-vegas-casino-hotel-market-outlook-2014/

Ooms, J., Lang, D. T., & Hilaiel, L. (2015). Jsonlite: A Robust, High Performance JSON Parser and Generator for R (Versão 0.9.17). Acedido outubro, 1, 2015, em https://cran.r-project.org/web/packages/jsonlite/index.html.

Paltoglou, G., & Thelwall, M. (2012). Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media. ACM Trans. Intell. Syst. Technol., 3(4), 66:1-66:19. Acedido janeiro 12, 2015, em http://doi.org/10.1145/2337542.2337551.

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

Parameswaran, M., & Whinston, A. B. (2007). Research issues in social computing. *Journal of the Association for Information Systems, 8*(6), 336-350.

Pekar, V., & Ou, S. (2008). Discovery of subjective evaluations of product features in hotel reviews. *Journal of Vacation Marketing, 14*, 145–155.

Piatetsky, G. (ed.), (2014). R leads RapidMiner, Python catches up, Big Data tools grow, Spark ignites. Acedido junho 11, 2015, em http://www.kdnuggets.com/2015/05/poll-r-rapidminer-python-big-data-spark.html.

Piatetsky, G. (ed.), (2015). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Acedido junho 11, 2015, em http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130-137.

Power, D. J., & Phillips-Wren, G. (2011). Impact of social media and Web 2.0 on decision-making. *Journal of decision systems, 20*(3), 249-261.

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics, 3*(2) 143-157.

Provost, F., Fawcett, T., (2013). *Data science for business.* Sebastopol. O'Reilly.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers: San Mateo, USA. ISBN: 1-55860-238-0.

Sharda, R., Delen, D., Turban, E., (2014). *Business intelligence: a managerial perspective on analytics.* (3rd ed.). Harlow. Pearson.

Stieglitz, S., & Dang-Xuan, L. (2012). Political Communication and Influence through Microblogging - An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior (pp. 3500–3509). IEEE. Acedido junho 11, 2015, em http://doi.org/10.1109/HICSS.2012.476.

Stitson, M. O., Weston, J. A. E., Gammerman, A., Vovk, V., & Vapnik, V. (1996). Theory of support vector machines. University of London. Acedido outubro 18, 2015, em http://sites.google.com/site/jeisongutierrez/TheoryofSVMachines.pdf

Sun, J., Wang, G., Cheng, X., & Fu, Y. (2014). Mining affective text to improve social media item recommendation. *Information Processing & Management*. Acedido março 18, 2015, em http://doi.org/10.1016/j.ipm.2014.09.002.

Tan, A. (1999). Text Mining: The state of the art and the challenges. *In Proceedings of the PAKDD Workshop on Knowledge Discovery from Advanced Databases* (pp. 65-70).

Tarannum, N., Rizvi, S.H., Keole, R.R. (2015). A Preliminary Review of Web-Page Recommendation in Information Retrieval Using Domain Knowledge and Web Usage Mining. *International Journal of Advance Research in Computer Science and Management Studies.* Volume 3, Issue 1. ISSN: 2321 – 7782. Research Article / Survey Paper / Case Study. Acedido fevereiro 5, 2015, em www.ijarcsms.com.

Thevenot, G. (2007). Blogging as a social media. *Tourism & Hospitality Research, 7(3/4),* 287-289.

Thiel, K., Kotter, T., Berthold, M., Silipo, R., Winters, P. (2012). Creating Usable Customer Intelligence from Social Media Data: Network Analytics meets Text Mining. Revision: 120403F. Acedido dezembro 15, 2015, em https://www.knime.org/files/knime_social_media_white_paper.pdf.

TripAdvisor. (2015). About TripAdvisor. Acedido agosto 10, 2015, em http://www.tripadvisor.com/PressCenter-c6-About_Us.html.

Trivago. (2015). Who we are. Acedido agosto 10, 2015, em http://www.trivago.com/static/company/company.

UNWTO. (2015a). Glossary of tourism terms. Acedido abril 18, 2015, em https://s3-eu-west-1.amazonaws.com/staticunwto/Statistics/Glossary+of+terms.pdf

UNWTO. (2015b). Who we are. Acedido abril 18, 2015, em http://www2.unwto.org/content/who-we-are.

UNWTO. (2015c). Why tourism. Acedido abril 18, 2015, em http://www2.unwto.org/content/why-tourism.

Vapnik, V., Golowich, S. E., & Smola, A. (1996). Support vector method for function approximation, regression estimation, and signal processing. In *Advances in Neural Information Processing Systems 9*. Citeseer. Acedido outubro 15, 2015, em http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.3139

Vapnik, V. (2006). Estimation of dependences based on empirical data. Springer Science & Business Media. Acedido outubro 15, 2015, em https://www.google.com/books?hl=pt-PT&lr=&id=N_-5VRWai84C&oi=fnd&pg=PA411&dq=Estimation+of+Dependences+Based+on+Empirical+Data&ots=ReEBtnkJR_&sig=VgnEs25-HpBDOntKzzvxmimSV8U.

Wang, G., Sun, J., Ma, J., Xu, K., & Gu, J. (2014). Sentiment classification: The contribution of ensemble learning. *Decission Support Systems, 57*, 77-93.

Xiang, Z., & Gretzel, U. (2010). Role of social media in online travel information search. *Tourism Management, 31*(2), 179–188. doi:10.1016/j.tourman.2009.02.016.

Yap, B.W., Rani, K. A., Rahman, H. A. A., Fong, S. Khairudin, Z., & Abdullah, N. N. (2014). An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. Em Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013) (pp. 13-22). Springer. Acedido outubro 1, 2015, em http://link.springer.com/chapter/10.1007/978-981-4585-18-7_2.

Yelp. (2015a). About us. Acedido abril 19, 2015, em http://www.yelp.com/about.

Yelp. (2015d). Recommended Reviews. Acedido abril 19, 2015, em http://www.yelp-support.com/Recommended_Reviews?l=en_US.

Zeng, B., Gerritsen, R. (2014). What do we know about social media in tourism? A review. *Tourism Management Perspectives. 10*, 27-36.