ANÁLISE DE SENTIMENTOS NA CLASSIFICAÇÃO DE ONLINE REVIEWS APLICANDO TEXT MINING

Sentiment Analysis in Online Reviews Classification using Text Mining Techniques

Águeda, M.*; Rita, P.**; Guerreiro, P.***
* Instituto Universitário de Lisboa (ISCTE-IUL)
** NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Portugal
*** Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal

Resumo — **O crescimento dos *sociais media* proporcionou nos últimos anos um aumento de comentários *online* que refletem as opiniões dos consumidores. As empresas beneficiam muito da disponibilização desta informação quer para darem uma resposta mais eficaz à insatisfação dos consumidores quer explorando oportunidades de mercado, observando os padrões que poderão representar necessidades insatisfeitas. O presente estudo tem como objetivo dar resposta a esta problemática através de uma pesquisa assente na plataforma Yelp. Para tal, foram extraídos 14.000 comentários, relacionados com diferentes produtos turísticos e aplicadas técnicas de *text mining* e *topic models* de modo a encontrar os principais tópicos latentes discutidos nos comentários online e os seus sentimentos. O estudo apresenta 20 tópicos latentes das discussões online e revela que o tópico que discute temas como *"Air Travel"* é o que obtém força de sentimento menos positiva em termos médios, devendo portanto ser alvo de uma análise mais profunda.**

Palavras Chave - ***Text Mining*; *Reviews Online; Análise de Sentimentos*; *Processamento de Linguagem Natural*; *Topic Models*.**

Abstract — **The growth of social media in recent years has led to an increase in online reviews that reflects consumer opinions. Firms benefit greatly from making this information available in order to respond more effectively to consumer dissatisfaction and to exploit market opportunities by observing standards that may represent unsatisfied needs. The present study aims to address this problem through a survey based on the Yelp platform. To this end, 14,000 comments related to different tourism products were used and text mining techniques and topic models were applied to find the main latent topics discussed in the online comments and their associated sentiments. The study presents 20 latent topics from online discussions and reveals that the topic that discusses "Air Travel" themes is the one with a lower sentiment connotation on average and should therefore be the subject of a deeper evaluation.**

Keywords **- Text Mining*; Online Reviews; Sentiments Analysis; Natural Language Processing; Topic Models.***

I. INTRODUÇÃO

Há cada vez mais consumidores com acesso à Internet que vêm ganhando interesse e dependência pelos serviços disponibilizados online. Esta interação no mundo virtual traduziu-se num grande volume de dados gerados diariamente, quer seja através das redes sociais, blogs, fóruns [1] 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 [2]. Os comentários online ganharam assim uma importância fundamental na decisão, quer dos consumidores quer das empresas. Do lado dos consumidores porque a recomendação de bens e serviços os ajuda 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. Do lado das empresas, porque estas podem acompanhar como os seus bens e serviços estão a ser apreciados pelos seus consumidores, o que se poderá traduzir-se no crescimento reputacional da marca e também na exploração de novas oportunidades de negócio.

Criar e executar este filtro que ajude quer consumidores quer empresas a tomar melhores decisões no meio da diversidade da informação online não é fácil, pois no meio de tantos dados surge sempre o problema de confiabilidade, fidedignidade e utilidade destas informações [3] que podem muitas vezes escapar aos olhos mais atentos. Para isso, é preciso garantir que as informações analisadas nos social media sejam, de fato, úteis para essa tomada de decisão. Empresas como a Yelp, a Booking e outros sites de recomendação disponibilizaram plataformas que permitem aos seus utilizadores publicarem as suas experiências. Essa informação é preciosa sobretudo se existir a capacidade de conhecer quais os principais elementos determinantes para a satisfação dos clientes (abordados de forma positiva, negativa ou neutra) de modo a poder identificar oportunidades no mercado.

O turismo é um mercado cada vez mais relevante na economia mundial e, apesar da importância da informação acerca do comportamento online dos consumidores, poucas são

as empresas que, de forma estruturada, recorrem a essa informação para fins de posicionamento. Apesar da investigação feita nesta área sugerir que é importante explorar em detalhe os padrões de comportamento dos consumidores online [4][5][6][7][8], é necessário uma abordagem que permita às organizações analisar a satisfação dos consumidores relativamente a tópicos de interesse latentes presentes nas opiniões positivas e negativas online e não focar a análise apenas em palavras isoladas que por vezes pouca relevância têm se não forem enquadradas num contexto mais geral.

As contribuições teóricas do presente artigo focam-se em dois aspetos. Primeiro, na apresentação de uma metodologia de análise de informação não estruturada que permite replicar esta pesquisa científica noutros contextos para extrair tópicos latentes das opiniões dos consumidores. Por outro lado, a investigação centra-se na indústria do turismo para analisar quais os tópicos latentes que apresentam maior satisfação.

Este artigo tem também importantes contribuições práticas, nomeadamente ao permitir às organizações compreender quais os temas pelos quais os clientes têm mais preferência podendo assim melhorar a sua aposta nesses segmentos, ou compreender aqueles segmentos que são menos satisfatórios para os clientes e que poderão representar uma potencial oportunidade de negócio ainda por explorar.

II. Revisão Biliográfica

A. *Social media e a sua influência na Indústria Turística*

Os social media vêm 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. 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 [10]. Uma forma de diminuir esta dificuldade é a aproximação das empresas do turismo ao consumidor através dos canais digitais, 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 perceção de qualidade dos produtos oferecidos.

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 dos turistas se sentirem assegurados, procurando produtos que já venham recomendados e bem referenciados por outras pessoas com experiência de utilização [10][11][12][13].

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 visitantes a tomarem as suas decisões [9]. 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 que, com frequência, um testemunho de um turista prévio possa ser bem mais poderoso do que a comunicação tradicional através de campanhas publicitárias [10].

No entanto o número de comentários e recomendações têm vindo a aumentar exponencialmente na última década e é necessário encontrar mecanismos eficazes de extração de padrões no meio da grande quantidade de informação disponível, salientando a mais relevante para a tomada de decisão das empresas. Este artigo utiliza uma técnica de extração de padrões dos comentários de texto (text mining) como meio de atingir esse objetivo.

B. *Text Mining*

Text Mining é 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 [14]. Na terminologia de text mining, um documento é um conjunto de caracteres em forma de texto, composta por tokens, conjuntos de unigramas (termos individuais) ou n-gramas (termos com mais do que uma palavra) que em conjunto representam o corpus [15][16].

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 destes tipos de dados que se encontram disponíveis dentro e fora das organizações, como os provenientes das plataformas web. O text mining utiliza-os para obter conhecimento capaz de resolver problemas do mundo real [17] e oferece inúmeros benefícios às organizações, principalmente para as que lidam diariamente com grandes quantidades de dados textuais.

C. *Processamento de Linguagem Natural (PLN)*

O processamento de linguagem natural é um elemento fundamental da análise de informação não estruturada uma vez que permite que sejam analisados não apenas tokens fora do seu contexto (bag of words), mas analisando o texto a um nível semântico bastante mais abstrato [18]. O PLN 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 [16].

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 (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 traduz o conhecimento implícito [19].

D. *Análise de Sentimentos*

Durante a última década, tem havido um interesse crescente na área do PLN, a base inerente à análise de sentimentos. Com várias nomenclaturas como emotional polarity analysis, review mining, subjectivity analysis, opinion mining e appraisal extraction [20][16], a análise de sentimentos é considerada um processo que deteta automaticamente o conteúdo emocional ou opinativo presente num texto e determina a sua polaridade [21]. A polaridade dos sentimentos é uma característica particular do texto, normalmente dicotomizado em positivo e negativo, ou por um intervalo de valores.

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 [16]. Mas de entre as muitas definições existentes, encontram-se algumas que vão mais ou menos de encontro à análise de dados e descoberta de conhecimento que fazem parte do âmbito do presente artigo. Mostafa [22] define-o como uma técnica de extração automática de conhecimento através de um léxico de sentimentos que classifica a conotação e o peso do sentimento em cada palavra.

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 [16] [23] ou até mesmo com um intervalo de força de opinião [1]. 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.

III. METODOLOGIA

A. *Amostra*

Para a análise presente neste artigo, foram recolhidos 14.000 reviews de forma aleatória de um conjunto de opiniões disponíveis num dataset da plataforma Yelp [24]. Os comentários incidem sobre as mais diversas empresas em redor das universidades de Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas e Madison nos Estados Unidos.

B. *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 [25]. Esta etapa normalmente diferencia os processos de text mining dos de data mining uma vez que num projeto de data mining os dados já se encontram estruturados.

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 tais como conversão para minúscula; remoção de números; remoção de pontuação; remoção de *stopwords* [25]; remoção de mais palavras consideradas irrelevantes para a análise e diretamente relacionadas com o assunto, por exemplo yelp, URL e yummy; transformação de alguns termos considerados importantes [26], por exemplo passar todos os termos bbq para barbecue; aplicação do processo de stemming [25] e remoção de espaços em branco. Aquando da remoção da pontuação detetou-se que algumas palavras ficavam coladas umas às outras. Isto porque depois do sinal de pontuação, em alguns casos, não existia espaço antes de começar a frase seguinte. Este facto produzia termos incorretos prejudicando a análise. Uma solução encontrada foi a de substituir as pontuações por um espaço em branco em vez de as remover. A remoção das *stopwords* foi conduzida através de um processo de remoção de verbos auxiliares, artigos, pronomes, preposições e interjeições que não são tipicamente incluídos numa análise de text mining [25]. O *stemming* é um processo de normalização linguística e reduziu as palavras aos seus radicais utilizando o algoritmo de Porter [27]. Neste caso, os prefixos e sufixos de cada termo foram eliminados através de um processo automático para que palavras como universally e universal, por exemplo, pudessem estar enquadradas no mesmo conceito [27] [25]. Posteriormente às transformações foi construído o Document-by-term matrix (DTM), uma matriz com as frequências absolutas dos vários termos pelos vários reviews, aceitando unigramas e bigramas não inferiores a três carateres [28].

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. Estes resultados são problemáticos uma vez que indicam 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) foi indicado um limite de termos entre 1 até 3 n-gramas, (2) estabeleceu-se que o tamanho mínimo do termo (minWordLength) a considerar de 4 caracteres, (3) a frequência mínima dos termos no documento (minDocFreq) de 2 e (4) foi novamente elaborada a fase de transformação com os métodos de stemming, remoção de stopwords e remoção de números. A DTM final revelou 625.851 termos em 14.000 comentários.

Geralmente, o desempenho dos algoritmos de reconhecimento de padrões é muito prejudicado com bases 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 [28]. 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 [28]. Isto levou a que fosse aplicada a medida term frequency inverse document frequency (TF-IDF) [29] para reduzir a dispersão. Esta combina a frequência do termo com o inverso da frequência do

documento, salientando termos com alta frequência no 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 corpus. Com a aplicação desta técnica apenas os termos com frequência superior à mediana [30] continuaram a fazer parte da DTM. Depois deste processo a DTM ficou reduzida a 13.917 documentos para 339.928 termos.

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, a wordcloud manteve o foco nos termos mais relacionados com esta área. Tendo terminado a análise sobre a DTM, o próximo passo consistiu em 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, no entanto mais recentemente têm sido utilizados com sucesso métodos de mixed-clustering como os topic models, onde os termos se caracterizam por uma multi-pertença a tópicos latentes, o que é mais indicado para o caso da análise textual [31] [30] [29] [32]. Os topic models são modelos probabilísticos que permitem descobrir a estrutura semântica subjacente à coleção de documentos tendo por base modelos Bayesianos que atribuem a cada termo uma probabilidade de pertença a vários tópicos latentes [33].

Embora existam vários algoritmos de topic models, o correlated topic model (CTM), baseado no algoritmo latent dirichlet allocation (LDA) tem sido utilizado com sucesso no tratamento de texto [29]. Em comparação com o seu antecessor que utilizava tópicos do tipo bag-of-words, o CTM assume uma correlação entre os tópicos latentes [29] e é mais eficaz em termos da sua perplexity, uma medida de eficácia que indica a precisão do modelo a prever as palavras de um documento após observar apenas uma parte deste.

Foram testados modelos potenciais com clusters que variaram entre modelos com apenas 2 tópicos e modelos com 60 tópicos, e foi avaliada a sua log-likelihood e perplexity, O número tópicos foi determinado quando a variabilidade explicada não se alterou significativamente ao aumentar o número de possíveis tópicos latentes [31]. A figura 1 apresenta o o log-likelihood e perplexity para os modelos testados.

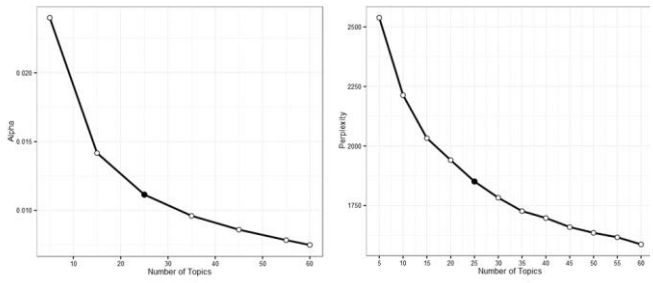

Figura 1 - Valores de Log-likelihood (Alpha) e Perplexity do CTM por número de tópicos

De acordo com a figura 1, 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 apresentadas no site da Yelp. Uma análise de sensibilidade a cada um dos tópicos indica que a amostra com os tópicos gerados pelo algoritmo de topic models representa predominantemente 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. No caso da categoria Shopping ser 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. No entanto, com a análise de alguns comentários relacionados com cada um dos tópicos foi possível chegar às nomenclaturas apresentadas na Tabela 1. Esta tabela apresenta os 20 tópicos e os cinco termos mais correlacionados com cada um, bem como as respetivas designações.

TABELA 1 – TÓPICOS

1 - Buffet	**2 - American Restaurant**	**3 - Hotel**	**4 - Mexican Restaurant**	**5 - Pizza Restaurant**
941 rev.	**856 rev.**	**786 rev.**	**760 rev.**	**760 rev.**
Buffet	Burger	Pool	Taco	Pizza
Breakfast	Coffee	Show	Salsa	Crust
Cake	Ice cream	Burger	Steak	Store
Roll	Breakfast	Store	Store	Burger
Pasta	Sandwich	Customer	Burger	Sandwich
6 - Shopping Center	**7 - Lounge Area**	**8 - Japanese Restaurant**	**9 - Restaurant and Nails Salon**	**10 - Italian Restaurant**
739 rev.	**730 rev.**	**729 rev.**	**705 rev.**	**682 rev.**
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 rev.	**678 rev.**	**658 rev.**	**657 rev.**	**652 rev.**
Sandwich	Movie	Show	Flight	Class
Crepe	Pool	Burger	Store	Pizza
Store	Theater	Store	Airport	Burrito
Customer	Show	Donut	Noodle	Tour
Show	Store	Wing	Show	Noodle
16 - Meeting Place	**17 - Casino**	**18 - Fast Food Restaurant**	**19 - Steak House**	**20 - Food**

645 rev.	644 rev.	630 rev.	627 rev.	353 rev.
Store	Game	Burger	Burger	Food price
Burger	Pizza	Chili	Barbecue	Food service
Buffet	Burger	Sandwich	Coffee	Food food
Mall	Store	Pizza	Sandwich	Food
Game	Wing	Coffee	Chop	Folk

Na preparação para a análise de sentimentos foram adicionados ao software *Lexalytics Semantria* os tópicos construídos com o CTM de forma a categorizar o sentimento dos tópicos latentes. Os 20 tópicos latentes foram classificados tendo em conta a sua polaridade (positivo, negativo, neutro), analisando a força do sentimento dos termos que compõe cada tópico. O *Lexalytics Semantria* classifica os tópicos latentes tendo em conta os sentimentos dos termos que forem encontrados nas reviews e que façam parte de cada tópico, numa escala logarítmica que varia entre -10 e 10. Cada review é classificada como tendo presente ou não o tópico latente e para cada tópico é gerado um valor contínuo que representa a força desse sentimento.

TABELA 2 – CLASSIFICAÇÃO DE SENTIMENTOS DOS TÓPICOS LATENTES

Tópico Latente	**Negativo**	**Neutro**	**Positivo**	**# Reviews**
Air Travel	10	138	31	179
American restaurante	46	692	330	1068
Buffet	82	1282	687	2051
Casino	36	543	279	858
Entertainment	33	370	183	586
Fast food restaurante	192	2893	1504	4589
Food	238	3165	1836	5239
French Bakery	168	1342	794	2304
High Class Place	7	204	90	301
Hotel	175	1407	831	2413
Italian Restaurant	133	2180	1137	3450
Japanese restaurante	86	1209	712	2007
Lounge área	60	835	444	1339
Meeting place	37	464	190	691
Mexican restaurante	102	1550	778	2430
Music Venue	46	623	301	970
Pizza restaurante	97	1506	815	2418
Restaurant and Nails Salon	36	641	383	1060
Shopping Center	43	543	233	819
Steak house	132	1882	956	2970
Total	**1759**	**23469**	**12514**	**37742**

A tabela 2 mostra o número de vezes que o tópico latente é abordado nos reviews classificados como negativo, neutro ou positivo.

A tabela 2 mostra que nos tópicos analisados, a maioria das discussões foram classificadas com o sentimento neutro (23.469), havendo muito mais discussões classificadas como positivas (12.514) do que como negativas (1.759).

No entanto, foi elaborada uma análise da média dos sentimentos em cada tópico latente (figura 2) de forma a verificar quais os tópicos em que, independentemente do número de reviews, possa haver uma maior ou menor força de sentimento expresso em cada review.

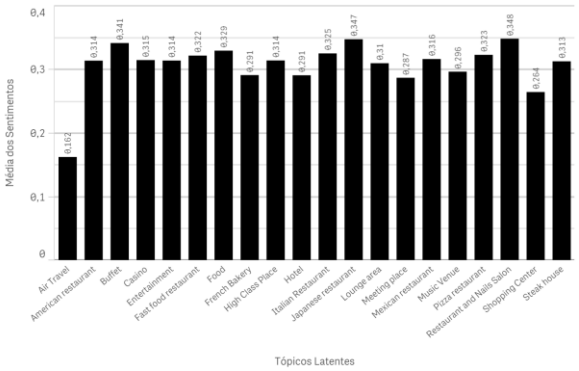

Figura 2 - Média de Sentimentos por Tópicos

Como se pode verificar na figura 2, o tópico latente com a média de sentimentos mais reduzida é o tópico Air Travel, o que significa que embora existam mais sentimentos positivos do que negativos (como revelou a tabela 2), os reviews que abordam este tópico de forma negativa têm uma conotação mais forte que não permite que este tópico tenha um sentimento médio tão elevado como os restantes. Outro facto a salientar é que um dos tópicos com menor número de reviews a referenciá-lo, High Class Place (301) não foi um dos que obteve menores classificações em termos de média dos sentimentos, o que indica que apesar de tudo os comentários positivos são mais fortes em termos de sentimento do que os comentários negativos. Este resultado condiz com o tipo de tópico identificado que aponta para experiências de elevado valor. A análise revela também que os tópicos que agregam termos relativos a "Restaurants and Nails Salon" e "Japanese Restaurant" são os que têm um conjunto de comentários com uma conotação positiva mais elevada.

IV. CONCLUSÕES

Este artigo apresenta uma análise exploratória de text mining com o objetivo de encontrar padrões que revelem quais os principais tópicos latentes abordados pelos utilizadores e o seu sentimento nas plataformas online, de forma a poder ajudar as empresas a tomar melhores decisões de negócio relativamente a oportunidades que o mercado possa oferecer. O electronic word-of-mouth (e-WOM) é hoje um veículo fundamental que as empresas não podem ignorar, nomeadamente auscultando as opiniões dos consumidores acerca das suas experiências.

O Yelp permite aos utilizadores discutir as suas experiências através dos seus comentários online. Um tópico abordado de forma negativa ou positiva pode representar uma importante oportunidade de negócio através da exploração da diferenciação das empresas que se querem afirmar no mercado. Por exemplo, o presente artigo apresenta resultados que indicam que o tópico "Air Travel" é frequentemente abordado com um sentimento mais negativo do que os restantes tópicos latentes nas discussões. Esta análise permite tirar conclusões e explorar melhor quais os drivers que poderão estar por detrás deste padrão de forma a que (1) as empresas com este tipo de experiências possam corrigir esta insatisfação e (2) as empresas que querem encontrar oportunidades de mercado possam explorar esta insatisfação indo de encontro às reais necessidades dos consumidores.

As limitações deste estudo focam-se sobretudo na utilização de uma única plataforma de recomendações online como base amostral. Futuras pesquisas podem utilizar a metodologia aqui apresentada para confirmar a existência de padrões de sentimento noutros contextos.

REFERÊNCIAS BIBLIOGRÁFICA

[1] B. Pang, and L. Lee, L. "Opinion mining and sentiment analysis. foundations and trends in information retrieval", vol. 2/1-2, pp. 1–135, 2008.

[2] P. Greenberg, "The impact of CRM 2.0 on customer insight." The Journal of Business and Industrial Marketing, vol. 25/6, pp. 410–419, 2010.

[3] P. Hajas. L. Gutierrez, and M.S. Krishnamoorthy, "Analysis of Yelp Reviews", Recuperado de https://www.researchgate.net/publication/263736290_Analysis_of_Yelp_Reviews, 2014.

[4] M. Dwivedi, T.P. Shibu, and U. Venkatesh, "Social software practices on the Internet: implications for the hotel industry", International Journal of Contemporary Hospitality Management, vol. 19/5, pp. 415-426, 2007.

[5] G. Thevenot, "Blogging as a social media", Tourism & Hospitality Research, vol. 7/3/, pp. 287-289, 2007.

[6] S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. Reis and J. Reynar, "Building a sentiment summarizer for local service reviews," in WWW Workshop on NLP in the Information Explosion Era, vol.14, 2008.

[7] V. Pekar and S. Ou, S., "Discovery of subjective evaluations of product features in hotel reviews", Journal of Vacation Marketing, vol. 14, pp. 145–155, 2008.

[8] I.S. Abeywardena, "Public opinion on OER and MOCC: A sentiment analysis of twitter data", in International conference on open and flexible education (ICOFE 2014). Hong Kong, China, 2014.

[9] B. Zeng, R. Gerritsen, "What do we know about social media in tourism? A review", Tourism Management Perspectives, vol.10, pp. 27-36, 2014.

[10] S.W. Litvin, R.E Goldsmith, and B. Pan, "Electronic word-of-mouth in hospitality and tourism management", Tourism Management, vol. 29/3, pp. 458-468, 2008.

[11] D. Leung, R. Law, H. van Hoof and D. Buhalis, "Social media in tourism and hospitality: a literature review", Journal of Travel & Tourism Marketing, vol. 30/1, pp. 3–22, 2013.

[12] J. Fotis, "Discussion of the impacts of social media in leisure tourism: the impact of social media on consumer behavior: Focus on leisure travel". Recuperado de http://johnfotis.blogspot.com.au/2012_03_01_archive.html, 2012.

[13] A.M. Munar, and J.K.S. Jacobsen, "Motivations for sharing tourism experiences through social media", Tourism Management, vol. 43, pp. 46–54, 2014.

[14] T.W. Miller, Data and text mining: a business applications approach, Upper Saddle River, NJ: Pearson Education International, 2005.

[15] F. Provost, T. Fawcett, Data science for business, Sebastopol. O'Reilly, 2013.

[16] R. Sharda, D. Delen and E. Turban, Business Intelligence and Analytics: Systems for Decision Support (10th edition). Pearson, 2015.

[17] S, Godbole, I. Bhattacharya, A. Gupta, and A. Verma, "Building Re-usable Dictionary Repositories for Real-world Text Mining", in Proceedings of the 19th ACM International Conference on Information and Knowledge Management, New York, NY, USA: ACM, 2010, pp. 1189–1198.

[18] E. Cambria, and B. White, Jumping "NLP curves: a review of natural language processing research [review article]", IEEE Computational Intelligence Magazine, vol. 9/2, pp. 48-57, 2014.

[19] S. Feldman, NLP meets the jabberwocky: natural language processing in information retrieval: Search Engine Section, (Online, Ed.). Information Today, 1999.

[20] B. Liu, Sentiment analysis and opinion mining. Morgan & Claypoool Publishers, 2012.

[21] G. Paltoglou and M. Thelwall, Twitter, "MySpace, Digg: Unsupervised Sentiment Analysis in Social Media", ACM Transactions on Intelligent Systems and Technology, vol. 3/4, pp. 1-19, 2012.

[22] M.M. Mostafa, "More than words: Social networks' text mining for consumer brand sentiments", Expert Systems with Applications, vol. 40, pp. 4241-4251, 2013.

[23] R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach", Journal of Informetrics, vol. 3/2, pp. 143-157, 2009.

[24] Yelp, Yelp Platform. Recuperado de https://www.yelp.com/, 2017.

[25] B, Liu, Web data mining: exploring hyperlinks, contents, and usage data. (2nd ed.). Springer Berlin Heidelberg New York, 2008.

[26] J. Bollen, A. Pepe and H. Mao, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena", in Fifth International Conference on Weblogs and Social Media, ICWSM, 2009, pp. 450–453.

[27] M.F. Porter, "An algorithm for suffix stripping", Program, vol. 14/3, 130-137, 1980.

[28] I. Feinerer, K. Hornik and D. Meyer, "Text mining infrastruture in R", Journal of Statistical Software. vol. 25/5, 2008.

[29] D.M. Blei and J.D. Lafferty, "A correlated topic model of science", The Annals of Applied Statistics, vol.1/1, pp. 17–35, 2007.

[30] B. Grün and K. Hornik, "topicmodels: An R Package for Fitting Topic Models", Journal of Statistical Software, vol. 40/13, 2011.

[31] J. Guerreiro, P. Rita and D. Trigueiros, "A Text Mining-Based Review of Cause-Related Marketing Literature", Journal of Business Ethics, vol. 139/1, pp. 1–18, 2016.

[32] T, Griffiths, and M. Steyvers, "Finding scientific topics", in Proceedings of the National Academy of Sciences of the United States of America, 2004, pp. 5228–5235.

[33] D.M. Blei, and J.D. Lafferty, "Topic models. text mining: classification, clustering, and application", vol. 10/71, pp. 34, 2009.