Tiradentes no TripAdvisor - O que se fala sobre essa simpática cidade histórica?

Antônio P. S. Alves[3], Lucas G. da Silva Félix[2], Carlos M. G. Barbosa[1], Vinícius da Fonseca Vieira[1] , Carolina Ribeiro Xavier[1]

[1]Departamento de Ciência da Computação
Universidade Federal de São João del Rei (UFSJ)
São João del-Rei - MG- Brasil

[2]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte- MG - Brasil

[3]Departamento de Ciência da Computação
Pontifícia Universidade Católica do Rio de Janeiro (PUC-RJ)
Rio de Janeiro - RJ - Brasil

carolinaxavie@ufsj.edu.br

Resumo. *O turismo é uma área que teve grandes impactos com expansão da internet. Hoje é possível planejar uma viagem de casa, usando somente informações da web. No entanto, os usuários chegaram a um ponto em que a quantidade de dados fornecidos pode ser mais confusa do que esclarecedora, causando um problema chamado de sobrecarga de informações. Assim, este trabalho se concentra reduzir este problema. Para isso, utiliza técnicas de mineração de texto para sumarizar as opiniões dos usuários e para entender o quão próximo ela está da classificação discreta dada por eles para um lugar ou atração no TripAdvisor. Como estudo de caso escolhemos a cidade de Tiradentes, localizada no interior de Minas Gerais, Brasil.*

Abstract. *Tourism is a field that had a significant impact with the expansion of the internet. Today it is possible to plan a trip from home, using only information from the web. Unfortunately, users have reached a point where the amount of data provided can be more confusing than the source of information, causing the overhead information problem. Thus, this work focuses on reducing this problem. For this, it uses text mining techniques to summarize user opinions on topics and to understand how close it is to the discrete rating given by them to a place or well on TripAdvisor. As a case study, we chose the city of Tiradentes, located in the interior of Minas Gerais, Brazil.*

1. Introdução

Por muitos anos, planejar uma viagem era sinônimo de comprar mapas e revistas especializadas sobre o destino, e passar horas decidindo qual trajeto realizar, quais atrações visitar e onde se hospedar. Esta situação mudou com o surgimento da Internet Colaborativa, que permitiu a geração de conteúdo por usuários e o compartilhamento de informações. Essa expansão, juntamente com o aumento do volume de dados gerados por usuários

nesse espaço, permitiu que o comportamento dos mesmos pudesse ser influenciado por conteúdos direcionados [Miguéns et al. 2008].

Nesse sentido, a migração de clientes e empresas para a web e o aumento no volume de dados tornaram a busca por informação uma tarefa desafiadora para os usuários. Não poderia ser diferente para os turistas, que possuem a sua disposição milhares de sites e blogs especializados em turismo, bem como redes sociais voltadas para este setor com milhões de opiniões sobre estabelecimentos disponibilizadas por viajantes de todo o mundo, como o Foursquare[1], Yelp[2] e Tripadvisor[3]. Segundo [Zheng et al. 2018], tamanha variedade de opções dificulta a tarefa de descobrir boas opções, causando assim um problema de sobrecarga de informações.

O objetivo principal deste trabalho é a coleta e a extração de informações para uma análise da consistência dos *reviews* e a sumarização das impressões dos turistas de atrações, restaurantes e hotéis de uma cidade. Na metodologia proposta são utilizadas duas técnicas de mineração de texto: Modelagem de Tópicos e Análise de Sentimento. Enquanto a primeira técnica foca na identificação de um conjunto de palavras que melhor sumariza diferentes documentos, a segunda possibilita a avaliação do sentimento presente em *corpus* de textos. A utilização destas duas técnicas é conhecida como Análise de Sentimento Baseada em Aspectos [Schouten and Frasincar 2016].

Utilizamos neste estudo a cidade de Tiradentes-MG, uma cidade histórica cuja economia é voltada para o turismo [Silveira 2014]. Por meio de nossa metodologia foi possível identificar as principais impressões dos turistas sobre a cidade, atrações e estabelecimentos. Foi estudado também algumas discrepâncias entre as avaliações em estrelas e o sentimento agregado ao *review* escrito pelo usuário. Localidades podem se beneficiar de nosso trabalho para identificar o que atrai e o que afasta os viajantes, enquanto os usuários podem utilizar os resultados alcançados para identificar locais que os agradam mais, dadas as características obtidas nos tópicos.

2. Fundamentação teórica

Modelagem de Tópicos (MT) consiste na tarefa da descoberta de relações entre Documentos e Tópicos, levando em consideração os Termos. Este tipo de técnica permite a sumarização, classificação, recuperação, organização e análise de documentos [Hofmann 1999]. Assim, cada tópico é constituído pela distribuição probabilística de palavras que co-ocorrem com frequência nos documentos. Atualmente, a MT tem sido vastamente aplicada na avaliação de grandes quantidades de texto, tornando possível a sumarização de documentos por meio de tópicos ou aspectos.

Na literatura, as técnicas de MT são divididas em duas abordagens: supervisionadas e não-supervisionadas. Em abordagens supervisionadas os modelos atuam com conhecimento prévio sobre os dados preditos [Ghosh et al. 2017], baseando-se em um conjunto de palavras que devem ser procuradas nos documentos. As abordagens não-supervisionadas, são as mais utilizadas na literatura, e trabalham para agrupar os dados de acordo com alguma medida de qualidade, sem qualquer conhecimento prévio. Neste

[1]https://pt.foursquare.com/
[2]https://www.yelp.com/
[3]https://www.tripadvisor.com.br/

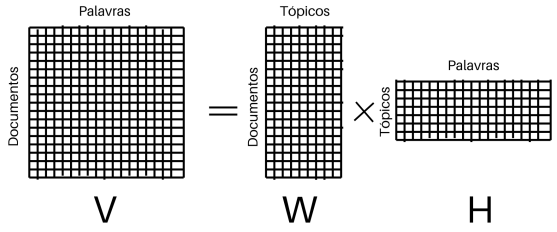

Figura 1. Modo de Funcionamento do NMF.

trabalho, utilizamos uma abordagem do segundo tipo, a Non-Negative Matrix Factorization (NMF)[Lee and Seung 2000].

O NMF realiza uma fatoração de matrizes, cujas matrizes resultantes não possuem valores negativos. Essa técnica depende somente da definição do número de tópicos e decompõe a matriz V (de documentos por termo) em duas outras matrizes, W (de documentos por tópicos) e H (de tópicos por termos), como mostra a Figura 1. Matematicamente, o NMF consiste na fatoração de V em uma matriz R, cuja combinação linear de W e H, aproxima-se de V [Eggert and Korner 2004], isto é, $V \approx R(W, H) = WH$.

As matrizes decompostas são de grande valia para análises posteriores, o que permite realizar uma análise de sentimento sobre os tópicos obtidos dos *reviews*, como em [Luiz et al. 2018] ou avaliar a agenda política e as temáticas dos discursos de parlamentares europeus através do tempo (1999-2014) [Greene and Cross 2016].

Vale ressaltar que apesar desta abordagem ser vastamente utilizada na literatura sua saída pode ser pouco interpretável, deixando dúvidas sobre a qualidade e congruência dos tópicos. Por este fato, diferentes trabalhos utilizam uma técnica chamada Informação Mútua Pontual Normalizada [Nguyen et al. 2018] (Normalized Pointwise Mutual Information - NPMI) , para avaliar a coerência dos tópicos gerados e consequentemente do modelo.

A Análise de Sentimento (AS) consiste na tarefa da detecção automática de polaridade (positiva, neutra, negativa) sobre o *corpus* do texto. Este tipo de técnica tem sido aplicada em diversos trabalhos de análise de redes sociais, permitindo a pesquisadores entender o comportamento de usuários, realizar avaliações de homofilia e avaliar opiniões em texto. Entender a conexão entre o sentimento dos usuários e as suas relações sociais tem um valor imenso para pesquisadores, indústria e, neste contexto, para o turismo [Yuan et al. 2014].

Encontramos na literatura a aplicação de três tipos principais de técnicas em AS: Modelos Baseados em Aprendizado de Máquina (AM), que treina modelos de AM sobre documentos rotulados e prediz instâncias não rotuladas; Baseado em Léxico, que define um dicionário léxico com o sentimento de polaridade de diversas palavras, utilizando regras gramaticais da linguagem para definir a polaridade de um texto; e Análise de Sentimento Baseada em Aspectos, que identifica palavras chave em documentos e mede o sentimento destas palavras.

3. Trabalhos relacionados

No contexto do turismo, AS tem sido vastamente aplicada para compreender a opinião de usuários em diversas atrações, hotéis e restaurantes [Valdivia et al. 2017], permitindo que estes possam compreender os tópicos que formam a opinião geral dos turistas [Chang et al. 2019]. Neste trabalho é considerada a AS baseada em léxico, muito utilizada na literatura [Luiz et al. 2018] e barata computacionalmente. A proposta de [Luiz et al. 2018] utiliza dados de aplicativos móveis para identificar aspectos considerados positivos e negativos sobre essas aplicações. O trabalho utiliza-se da técnica de NMF para identificação dos tópicos e da abordagem proposta em [Rocha et al. 2015] para análise de sentimento sobre os mesmos.

Na proposta de [Afzaal and Usman 2015], os autores utilizam dados de Londres com uma proposta básica para identificação de tópicos, já que identificam quais são os aspectos mais importantes através da frequência de substantivos nos documentos. Já o trabalho de [Afzaal et al. 2019] os autores focam em uma aplicação móvel que identifica os melhores locais em uma cidade através da identificação do sentimento sobre aspectos em *reviews* de locais. Para a identificação de tópicos é utilizado um conjunto de palavras presentes no léxico proposto por [Manning et al. 2014]. Após esse filtro de palavras, são identificadas aquelas que são sinônimos, para então calcular a frequência dos aspectos. Por meio dessa frequência, estes aspectos são então separados em grupos maiores que descrevem, de maneira mais genérica, os tópicos identificados.

4. Materiais e métodos

4.1. Dados utilizados

Neste trabalho utilizaremos os dados referentes à cidade de Tiradentes-MG, pelo fato de ser uma cidade conhecida pelos autores, devido a proximidade com a UFSJ e pelo fato de que boa parte de sua economia esta ligada ao turismo. Tiradentes possui um alto volume de dados produzidos por turistas na plataforma *TripAdvisor*, agregando 36 atrações, 171 hotéis, 147 restaurantes, com mais de 55000 *reviews* escritos por quase 35000 usuários, o que também justifica sua escolha.

Hoje em dia as páginas web não são mais estáticas, por tanto foi necessário utilizar um simulador de navegador, o qual, a partir de um link de uma página web, simula a navegação de um usuário, através de comandos pré-definidos e automáticos, como *scroll* e clique. Com este navegador e com a definição de comandos específicos, é possível *crawlear* os dados desejados em tempo de execução. Nosso *crawler* foi construído utilizando as bibliotecas **Selenium**[4], para a simulação do navegador, e **BeautifulSoup**[5], para a recuperação e manipulação do *HTML* da página *web*. O uso conjunto dessas bibliotecas permitiu a extração e armazenamento dos dados selecionados referentes a cidade de Tiradentes, que datam de junho de 2008 a outubro de 2020.

4.2. Modelagem de Tópicos

Nesta etapa do trabalho, utilizamos a Modelagem de Tópicos (MT) para realizar uma sumarização dos *reviews* de usuários do *TripAdvisor*. Para esse fim, utilizamos a técnica de NMF.

[4]https://www.selenium.dev/

[5]https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Para representação das palavras é utilizada uma abordagem numérica baseada na matriz de *reviews* por palavras, conhecida na literatura como Frequência do Termo - Frequência de documentos inversa (*Term Frequency-Inverse Document Frequency (TF-IDF)*). Contudo, apenas com essa representação e aplicando o NMF, não é possível realizar um processo de extração de tópicos eficiente. Isso porque, a matriz completa de *reviews* por palavras, considerando todos os vocábulos que ocorrem, diminui a recuperação de termos mais importantes na clusterização de textos [Moh and Bhagvat 2012].

Desta maneira, antes de executar a MT, fez-se necessário a realização de um pré-processamento dos *reviews*. Apesar de não haver uma única sequência correta, genérica e funcional para bases textuais, foram estabelecidas três etapas, visando melhorar a coerência final dos tópicos resultantes. São elas: (i) Junção de nomes próprios, (ii) Remoção de entidades desnecessárias, (iii) Remoção de *stopwords* e outros componentes gramaticais que pouco agregam a compreensão.

A etapa desenvolvida em (i) é totalmente ligada ao conteúdo da base de *reviews*. Por ser uma cidade histórica e extremamente ligada ao catolicismo dos séculos XVIII e XIX, Tiradentes possui muitas atrações, hotéis e restaurantes que trazem consigo nomes de santos católicos. Portanto, visando evitar perder a referência para esses lugares - que podem ser importantes dentro do conjunto de documentos - estabelecemos esta etapa de pré-processamento, salvando o nome de santos que aparecem com um *underscore*. Assim, a Serra de São José, com esta etapa, é passada para as próximas etapas como "Serra de São_José". Esse processo difere do uso de bigramas e trigramas (n-gramas) pois só é feito sobre nomes de santos católicos. A aplicação de n-gramas sobre todo texto, piorou o resultado qualitativo obtido no final, por isso o filtro foi feito somente sobre os nomes de santos.

Em (ii), usamos um recurso muito comum na literatura, o de Entidade Nomeada Reconhecida (NER - Named Entity Recognition) [Gudivada and Arbabifard 2018], cujo objetivo é categorizar pessoas, lugares, organizações e outras entidades de interesse no texto. Com a utilização do NER foi possível avaliar que nos *reviews* realizados algumas entidades não agregavam valor para uma categorização textual. Essas entidades são: datas (exemplo: *visitei na* ***quarta-feira***), porcentagens (exemplo: *a atração está restrita a* ***10%*** *de sua capacidade em dias chuvosos*, números ordinais e cardinais (exemplo: *fui o* ***primeiro*** *a entrar em todos os* ***cinco*** *museus que visitei*) e pessoas (exemplo: *o* ***dono da pousada*** *era muito simpático*). As entidades identificadas dentro destas categorias foram descartadas, dado que as mesmas não agregam para a coesão dos tópicos gerados.

Por fim, em (iii) realizamos uma etapa comum a muitos trabalhos de mineração de texto, que é a remoção de *stopwords* e componentes gramaticais. *Stopwords*, diferente das entidades desconsideradas em ii são **palavras irrelevantes** para uma análise textual, como pronomes e preposições. Por não existir uma lista completa com estas palavras, utilizamos *stopwords* customizadas, baseadas na nossa *expertise* sobre a base e a lista de palavras utilizadas no artigo [Barbosa et al. 2019]. Todas as palavras contidas nessas listas são removidas. Além destas, também removemos os seguintes componentes gramaticais {pontuação, verbos, adjetivos, advérbios}, categorizados segundo o léxico definido no Universal Part-of-Speech tags (ou *Universal Part-of-Speech tags*[6]).

[6]https://universaldependencies.org/docs/u/pos/

Feito todo esse processo sobre a matriz de *reviews* por palavras, o NMF foi executado para ajustar o seu único parâmetro: o número de tópicos. O ajuste desse parâmetro traz matrizes resultantes diferentes, então o NPMI foi utilizado como métrica de avaliação da coerência dos tópicos obtidos nessas diferentes matrizes para obtenção do melhor número de tópicos. O número tópicos que maximiza o NPMI é 5, assumindo também o número de palavras que compõe cada tópico igual a 10.

4.3. Análise de Sentimento

A análise de sentimento (AS) é utilizada para identificar o sentimento nos tópicos extraídos na etapa anterior. Diferente do processo de MT, a AS não demanda um pré processamento de sua base de dados, tornando a tarefa mais simples. Entretanto, utilizamos um conjunto de dados cuja maioria dos *reviews* se encontram na língua portuguesa e grande parte das abordagens amplamente encontradas na literatura não possuem suporte ao português. Pensando nisto, fizemos uso da proposta [Almeida 2018], que utilizando como base a proposta de [Hutto and Gilbert 2014] desenvolveu um dicionário léxico para identificação de polaridades para língua portuguesa. A proposta dele, chamada Léxico para Inferência Adaptada (LeIA), propõe uma tradução do léxico do VADER.

O retorno dado pelo LeIA é uma quadrúpla $\{pst, ngt, nt, cmp\}$, que representa a positividade, negatividade, neutralidade e o *compound*, respectivamente, do texto lido. Sendo assim, fazemos uso somente do valor (cmp) - já que o mesmo representa uma média normalizada e ponderada do sentimento associado a um texto. Assim, associamos cada *review* a seu *compound*.

O *compound* de cada *review* é comparado à nota discreta dada pelo mesmo usuário que fez na plataforma, com o intuito de verificar se há uma discrepância entre a nota dada e o sentimento associado ao *review*. Para obtenção de valores compatíveis, fizemos a transformação da integral numérica gaussiana, que mapeia valores de -1 a 1, num intervalo genérico, conforme Equação 1.

$$x = \frac{1}{2}(min + max + cmp \times (max - min)) \tag{1}$$

Por meio dessa transformação foi possível associar o *compound* a uma das notas disponíveis no *TripAdvisor*, que são de 1 (muito negativo) a 5 (muito positivo). Desta maneira, após esta transformação, os valores de *compound* serão tabelados no intervalo.

5. Resultados

Considerando a discretização sobre o *compound* dos *reviews*, verificamos algumas peculiaridades nas avaliações. O comportamento dos usuários em relação às **Atrações**, como ilustra a Figura 2 mostra que temos uma pequena divergência: atrações com avaliações muito positivas (4 e 5 estrelas) têm valores de notas dadas superiores ao sentimento expresso nos *reviews*, e avaliações de 3 estrelas ou menos, possuem valores de sentimento superiores às notas dadas. Isso pode indicar que alguns usuários apresentam alguma inconsistência entre o que eles expressam nos textos dos *reviews* e o que atribuem nas notas em estrelas sobre as atrações, gerando um desequilíbrio que leva a falsa impressão de que uma atração é excelente (5 estrelas), mas na realidade ela poderia ser razoável (3 estrelas).

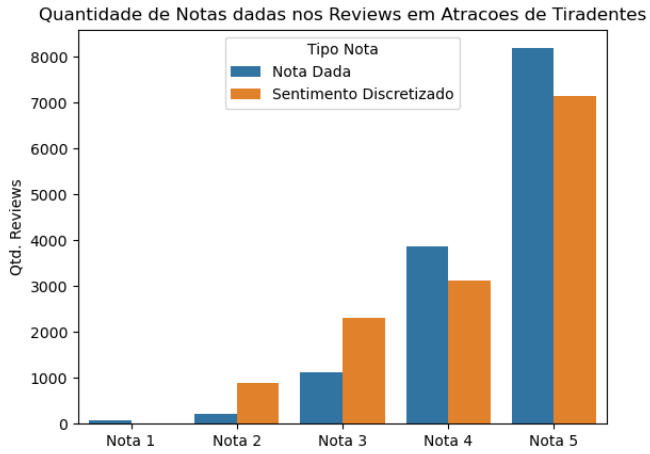

Figura 2. Sentimento Discretizado vs Notas Dadas pelos usuários em Atrações

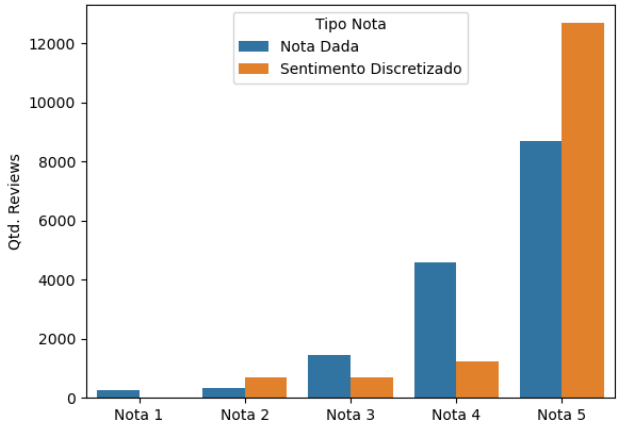

Figura 3. Sentimento Discretizado vs Notas Dadas pelos usuários em Hotéis

Em **Hotéis e Restaurantes**, Figuras 3 e 4 respectivamente, os usuários são mais exigentes para darem nota máxima (5 estrelas) - apesar de muitos terem expressado um sentimento muito positivo sobre estas localidades - e acabam dando mais notas de 4 e 3 estrelas. Isto explica a distribuição de notas dadas ser superior à distribuição do sentimento em notas inferiores.

Percebe-se que, nas Atrações, existem 5 tópicos bem definidos, referindo-se, respectivamente, à Igreja de Santo Antônio, ao trem Maria Fumaça que liga Tiradentes à São João del-Rei, ao Museu de Sant'Ana, à Serra de Tiradentes (Serra de São José) e ao centro histórico em geral. Percebemos que para todas essas atrações, o sentimento dos usuários é menor que a nota dada por eles - o que pode dificultar os responsáveis dessas atrações na identificação de pontos a serem melhorados de maneira direta, sendo necessário a leitura de inúmeros *reviews*. Podemos observar pelo terceiro valor que, a visita a Serra de Tiradentes e o centro histórico são os que apresentam a menor diferença entre o sentimento dos usuários e as notas dadas sobre estas atrações e poderia ser explicada pelo custo das mesmas, que diversas vezes é destacado como um preço acima da média (passeio de trem)

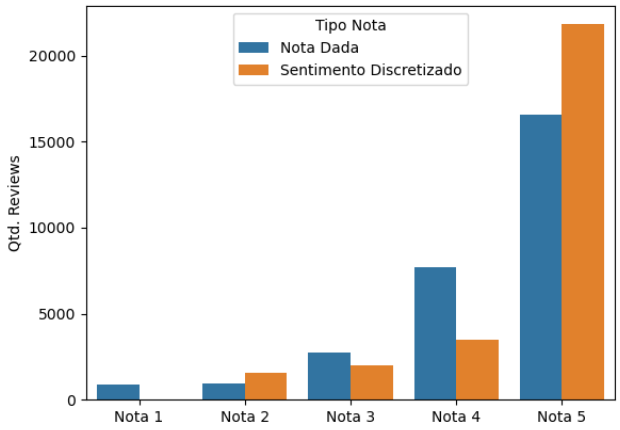

Figura 4. Sentimento Discretizado vs Notas Dadas pelos usuários em Restaurantes

Tópico	Resumo	Principais Palavras
1	Igrejas e Santuários	igreja; história; órgão; concerto; estilo barroco; santo antônio; arquitetura; riqueza; preço; altar
2	Passeios de Locomotiva	maria fumaça; passeio; são joão; preço passagem; crianças; passado; paisagem; história; trajeto; experiência
3	Museus Históricos	museu; cadeia; imagens; acervo; sant'ana ; peças; história; moderno; estrutura; coleção
4	Passeios Naturais	vista; serra; são josé; igreja; subida; trilha; visual; paisagem; caminhada; montanhas
5	Entretenimento Geral	centro histórico; ruas; histórias; bares; artesanato; conservado; lojinhas; passado; arquitetura; pedras

Tabela 1. Tópicos em Atrações de Tiradentes

Tópico	Resumo	Principais Palavras
1	Estadia	café manhã; piscina; vista serra; decoração; estadia; conforto; carinho; limpeza; acomodações; quartos
2	Comodidade	cama; banheiro; chuveiro; espaçoso; toalhas; limpeza; ventilador; travesseiros; colchão; frigobar
3	Atendimento	funcionários; atenciosos; simpático; prestativos; gentis; educados; solícitos; agradáveis; cordiais
4	Infraestrutura	localização; estacionamento; quartos; estrutura; piscina; centro histórico; bares; largo forras; lazer; crianças
5	Custo Benefício	custo beneficio; estacionamento; hospedagem; simples; localização; atrações; conforto; bares

Tabela 2. Tópicos em Hotéis de Tiradentes

Tópico	Resumo	Principais Palavras
1	Restaurantes Finos	carta vinhos; cardápio; opções; qualidade; localização; serviços; variedade; romântico; decoração; cervejas artesanais
2	Restaurantes Caseiros	comida mineira; qualidade; tipica; farta; feijão tropeiro; tutu; caseira; custo beneficio; tempero; rápido
3	Pizzarias	pizza; massa fina; pizzaria; crocante; sabores; recheio; qualidade; ingredientes; forno lenha; rápido
4	Bares	jantar; almoço; chopp; petiscos; cervejas; localização; experiência; cardápio; impecável; amigos; funcionários atenciosos
5	Lanchonetes	tiradentes; experiencia; sabor; hambúrguer; gastronomia; cozinha; comida; sorvete; goiabada; qualidade

Tabela 3. Tópicos em Restaurantes de Tiradentes

Tópico	SDM	MNE	RMSE
1	4,19	4,55	0,97
2	4,10	4,30	0,94
3	4,35	4,47	0,84
4	4,32	4,54	0,83
5	4,25	4,51	0,83

Tabela 4. Comparativo de Atrações de Tiradentes

ou como uma deselegância (existe uma crítica recorrente nas cidades históricas de Minas Gerais sobre a cobrança pra visitar o interior das igrejas). Mesmo assim, o sentimento discretizado associado aos tópicos, de maneira geral, são superiores à média de notas dadas. Podemos perceber que a visita à igrejas e o passeio de Maria fumaça é o que mais agrada nas atrações de Tiradentes.

Para os Hotéis, cada tópico traz um conjunto de atributos dos hotéis e pousadas: o tópico 1 é mais relacionado à estadia; o tópico 2, à infraestrutura dos quartos; tópico 3, ao atendimento; o tópico 4, à infraestrutura do hotel; e o tópico 5, ao custo benefício do mesmo. Nestes, a média das notas em estrelas é inferior ao sentimento discretizado médio encontrado nos *reviews*, podendo dar uma falsa impressão sobre a qualidade dos serviços prestados. É possível identificar, que *reviews* que tratam sobre a infraestrutura das acomodações e sobre o atendimento dos funcionários dos hotéis, apresentam menor discrepância entre o sentimento capturado e a nota em estrelas.

Por fim, na categoria dos Restaurantes, os tópicos encontrados são relacionados aos principais tipos de restaurantes existentes na cidade: o tópico 1, restaurantes mais finos; o tópico 2, restaurantes de comidas típicas; o tópico 3, pizzarias; o tópico 4, *pubs* e bares; e o tópico 5, ”restaurantes de rua”, tipo lanchonetes e docerias. Similarmente à situação dos hotéis, os usuários que avaliam restaurantes em Tiradentes dão notas mais baixas sobre o que eles expressam nos *reviews*. Nesta categoria podemos visualizar que os *reviews* sobre os *pubs* e ”restaurantes de rua”são mais consistentes, apresentando um erro médio menor entre as notas em estrelas e o sentimento capturado.

A categorização realizada pela MT sobre os *reviews* não informa o sentimento agregado. Essas questões são exploradas pela AS sobre os tópicos, em que é possível descobrir qual o sentimento médio associado à cada um dos tópicos, manipulando as matrizes resultantes do NMF. Nesse sentido, as Tabelas 4, 5 e 6 apresentam, para cada tópico obtido, uma comparação entre o Sentimento Discretizado Médio (SDM) definido pela Equação 1, a Média das Notas em Estrelas (MNE) e o *Root Mean Square Error* (RMSE) entre ambos. Ressalta-se que os valores possíveis para o SDM e a MNE estão entre 1 e 5, enquanto os valores possíveis para RMSE podem variar de 0 (SDM $\equiv$ MNE) até 4 (distância máxima entre o SDM e o MNE).

Percebe-se que para todas as atrações, o sentimento dos usuários é menor que a nota dada por eles - o que pode dificultar os responsáveis dessas atrações na identificação de pontos a serem melhorados de maneira direta, sendo necessário a leitura de inúmeros *reviews*. A partir de uma simples análise sobre o RMSE, percebe-se que a visita a Serra de Tiradentes e o centro histórico são os que apresentam a menor diferença entre o sentimento dos usuários e as notas dadas sobre estas atrações, indicando que seus visitantes são os mais coerentes.

Tópico	SDM	MNE	RMSE
1	4,66	4,38	0,97
2	4,74	4,16	0,87
3	4,78	4,52	0,67
4	4,70	4,33	0,78
5	4,71	4,32	0,79

Tabela 5. Comparativo de Hotéis de Tiradentes

Tópico	SDM	MNE	RMSE
1	4,58	4,42	0,93
2	4.46	4,22	0,85
3	4.47	4,1	0,82
4	4.77	4,3	0,79
5	4,57	4,22	0,79

Tabela 6. Comparativo de Restaurantes de Tiradentes

Nos Hotéis, a média das notas em estrelas é inferior ao sentimento discretizado médio encontrado nos *reviews*, podendo dar uma falsa impressão sobre a qualidade dos serviços prestados. É possível identificar, que *reviews* que tratam sobre a infraestrutura das acomodações e sobre o atendimento dos funcionários dos hotéis, apresentam menor discrepância entre o sentimento capturado e a nota em estrelas.

Na categoria dos Restaurantes, ocorre uma situação similar à dos hotéis, em que os usuários que avaliam restaurantes em Tiradentes dão notas mais baixas sobre o que eles expressam nos *reviews*. Nesta categoria, visualiza-se que os *reviews* sobre as bares e lanchonetes são mais consistentes, apresentando um erro médio menor entre as notas em estrelas e o sentimento capturado.

Por fim, para sumarizar todo o conteúdo processado nos comentários, geramos uma nuvem de palavras, que traz alguns termos interessantes. Podemos observar o destaque de alguns, como: localização, história, passado, centro histórico, arquitetura, paisagem, igreja, bares e qualidade. Que resumem de forma satisfatória muito do que a pequena cidade histórica oferece. Tiradentes é uma cidade que as pessoas costumam andar a pé, apreciar a paisagem, conhecer um pouco da história através da arquitetura, igrejas e museus, além de poder experimentar bares e restaurantes de diversos estilos.

Figura 5. Nuvem de palavras dos comentários das Atrações, Hotéis e Restaurantes de Tiradentes.

6. Conclusões e Trabalhos Futuros

Este trabalho permitiu avaliar o sentimento associado à Tiradentes, baseado nos *reviews* feitos por usuários na plataforma *Tripadvisor*. Para isso, utilizamos uma abordagem de AS Baseado em Léxico juntamente com MT para descobrir: quais são os principais tópicos relacionados às Atrações, aos Hotéis e aos Restaurantes e o sentimento associado a eles. Para esta etapa, utilizamos uma análise de sentimento baseada em léxico, enquanto que, para aquela etapa, utilizamos o NMF como estratégia de modelagem de tópicos. Os resultados obtidos nos permitiu verificar que enquanto as atrações de Tiradentes são levemente superestimadas, os restaurantes e hotéis são levemente subestimados, mas em geral, **a média das notas em estrelas dadas pode poupar usuários da leitura dos *reviews*, pois o erro médio em relação ao sentimento não excede uma estrela.** É importante levantar que existe algumas limitações na realização da comparação entre o sentimento expresso no texto aberto com a avaliação medida discretamente, em estrelas, duas delas são: o próprio método utilizado para realizar a análise de sentimentos e a forma que o usuário utiliza o campo aberto, que algumas vezes, pode ser usado para relatar exceções ao que foi avaliado em estrelas outras para confirmar ou justificar a nota dada.

Como trabalho futuro esperamos refinar as abordagens adotadas, experimentando outras técnicas de análise de sentimentos e realizando um estudo amostral de comentários discrepantes para entender o porquê da existência das diferença entre o que o usuário dá na nota discreta e o que ele escreve no texto livre.

Referências

[Afzaal and Usman 2015] Afzaal, M. and Usman, M. (2015). A novel framework for aspect-based opinion classification for tourist places. In *2015 Tenth International Conference on Digital Information Management (ICDIM)*, pages 1–9. IEEE.

[Afzaal et al. 2019] Afzaal, M., Usman, M., and Fong, A. (2019). Tourism mobile app with aspect-based sentiment classification framework for tourist reviews. *IEEE Transactions on Consumer Electronics*, 65(2):233–242.

[Almeida 2018] Almeida, R. J. A. (2018). Leia - léxico para inferência adaptada. `https://github.com/rafjaa/LeIA`.

[Barbosa et al. 2019] Barbosa, C. M. G., Felix, L. G. d. S., Xavier, C. R., and Vieira, V. d. F. (2019). A framework for the analysis of information propagation in social networks combining complex networks and text mining techniques. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web*, pages 401–408.

[Chang et al. 2019] Chang, Y.-C., Ku, C.-H., and Chen, C.-H. (2019). Social media analytics: Extracting and visualizing hilton hotel ratings and reviews from tripadvisor. *International Journal of Information Management*, 48:263–279.

[Eggert and Korner 2004] Eggert, J. and Korner, E. (2004). Sparse coding and nmf. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, volume 4, pages 2529–2533 vol.4.

[Ghosh et al. 2017] Ghosh, S., Chakraborty, P., Nsoesie, E. O., Cohn, E., Mekaru, S. R., Brownstein, J. S., and Ramakrishnan, N. (2017). Temporal topic modeling to assess associations between news trends and infectious disease outbreaks. *Scientific reports*, 7:40841.

[Greene and Cross 2016] Greene, D. and Cross, J. P. (2016). Exploring the political agenda of the european parliament using a dynamic topic modeling approach.

[Gudivada and Arbabifard 2018] Gudivada, V. N. and Arbabifard, K. (2018). Chapter 3 - open-source libraries, application frameworks, and workflow systems for nlp. In Gudivada, V. N. and Rao, C., editors, *Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications*, volume 38 of *Handbook of Statistics*, pages 31 – 50. Elsevier.

[Hofmann 1999] Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc.

[Hutto and Gilbert 2014] Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*.

[Lee and Seung 2000] Lee, D. D. and Seung, H. S. (2000). Algorithms for non-negative matrix factorization. NIPS'00, page 535–541, Cambridge, MA, USA. MIT Press.

[Luiz et al. 2018] Luiz, W., Viegas, F., Alencar, R., Mourão, F., Salles, T., Carvalho, D., Gonçalves, M. A., and Rocha, L. (2018). A feature-oriented sentiment rating for mobile app reviews. WWW '18.

[Manning et al. 2014] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

[Miguéns et al. 2008] Miguéns, J., Baggio, R., and Costa, C. (2008). Social media and tourism destination: Tripadvisor case study.

[Moh and Bhagvat 2012] Moh, T.-S. and Bhagvat, S. (2012). Clustering of technology tweets and the impact of stop words on clusters. In *Annual Southeast Regional Conference*, New York, NY, USA. Association for Computing Machinery.

[Nguyen et al. 2018] Nguyen, D. Q., Billingsley, R., Du, L., and Johnson, M. (2018). Improving topic models with latent feature word representations. *CoRR*, abs/1810.06306.

[Rocha et al. 2015] Rocha, L., Mourão, F., Silveira, T., Chaves, R., Sa, G., Teixeira, F., Vieira, R., and Ferreira, R. (2015). Saci: Sentiment analysis by collective inspection on social media content. *Web Semantics: Science, Services and Agents on the World Wide Web*, 34:27–39.

[Schouten and Frasincar 2016] Schouten, K. and Frasincar, F. (2016). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge & Data Engineering*, 28(03):813–830.

[Silveira 2014] Silveira, G. T. d. (2014). Turismo, emprego e renda: o caso da cidade histórica de tiradentes-mg.

[Valdivia et al. 2017] Valdivia, A., Luzón, M. V., and Herrera, F. (2017). Sentiment analysis in tripadvisor. *IEEE Intelligent Systems*, 32(4):72–77.

[Yuan et al. 2014] Yuan, G., Murukannaiah, P. K., Zhang, Z., and Singh, M. P. (2014). Exploiting sentiment homophily for link prediction. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 17–24.

[Zheng et al. 2018] Zheng, X., Luo, Y., Sun, L., Zhang, J., and Chen, F. (2018). A tourism destination recommender system using users' sentiment and temporal dynamics. *Journal of Intelligent Information Systems*, 51(3):557–578.