CENTRO UNIVERSITÁRIO LUTERANO DE PALMAS

Recredenciado pela Portaria Ministerial nº 1.162, de 13/10/16, D.O.U. nº 198, de 14/10/2016
AELBRA EDUCAÇÃO SUPERIOR - GRADUAÇÃO E PÓS-GRADUAÇÃO S.A.

Taylor Santos Oliveira

DESENVOLVIMENTO E VERIFICAÇÃO DO MÓDULO DE ANÁLISE DE SENTIMENTOS - NÍVEL DE DOCUMENTO DA SENTIMENTALL

Palmas - TO

2019

Taylor Santos Oliveira

DESENVOLVIMENTO E VERIFICAÇÃO DO MÓDULO DE ANÁLISE DE SENTIMENTOS - NÍVEL DE DOCUMENTO DA SENTIMENTALL

Projeto de Pesquisa elaborado e apresentado como requisito parcial para aprovação na disciplina de Trabalho de Conclusão de Curso II (TCC II) do curso de bacharel em Ciência da Computação pelo Centro Universitário Luterano de Palmas (CEULP/ULBRA).

Orientador: Prof. D.ra. Parcilene Fernandes de Brito.

Palmas - TO

2019

Taylor Santos Oliveira

DESENVOLVIMENTO E VERIFICAÇÃO DO MÓDULO DE ANÁLISE DE SENTIMENTOS - NÍVEL DE DOCUMENTO DA SENTIMENTALL

Projeto de Pesquisa elaborado e apresentado como requisito parcial para aprovação na disciplina de Trabalho de Conclusão de Curso II (TCC II) do curso de bacharel em Ciência da Computação pelo Centro Universitário Luterano de Palmas (CEULP/ULBRA).

Orientador: Prof. D.ra. Parcilene Fernandes de Brito.

Aprovado em: _____/_____/_____

BANCA EXAMINADORA

_____

Prof. D.ra Parcilene Fernandes de Brito

Orientador

Centro Universitário Luterano de Palmas – CEULP

_____

Prof. M.e Jackson Gomes de Souza

Centro Universitário Luterano de Palmas – CEULP

_____

Prof. M.e Fabiano Fagundes

Centro Universitário Luterano de Palmas – CEULP

Palmas - TO

2019

Dedico este trabalho a todas as pessoas que, devido aos motivos e dificuldades próprias de cada um, não tiveram a oportunidade ou foram impedidos de concluir uma graduação. Dedico também ao meu pai José Lopes de Oliveira e minha mãe Divanilda dos Santos Oliveira.

AGRADECIMENTOS

Agradeço primeiramente a Deus, que possibilitou que eu chegasse até aqui, me concedendo vida e saúde, suprindo todas as minhas necessidades. Também agradeço ao meu pai José Lopes de Oliveira e minha mãe Divanilda dos Santos Oliveira, que apesar das dificuldades financeiras em grande parte da minha caminhada até aqui, sempre se esforçaram com o intuito de possibilitar que eu tivesse o estudo que eles não tiveram a oportunidade. Também agradeço ao meu amigo Kalebe e meu amor Luennys Barbosa de Almeida, por seus incentivos e palavras de apoio, além de terem me ajudado com o processo da análise manual no processo de avaliação da qualidade do resultado do módulo desenvolvido neste trabalho.

Agradeço aos meus colegas de faculdade e todos os professores. Em especial a minha orientadora Parcilene, por ter aceitado o desafio de me orientar neste trabalho mesmo sem ter horários disponíveis. Foram alguns finais de semanas e horas além da conta me direcionando e ajudando. Sem ela certamente este trabalho não poderia atingir o resultado obtido. De forma especial, também quero agradecer aos professores Fabiano Fagundes e Jackson Gomes que fizeram parte da banca avaliadora desse trabalho. Nesse sentido, agradeço a todos que direta ou indiretamente contribuíram para a realização desse trabalho. A todos o meu Muito Obrigado!

“Só se pode alcançar um grande êxito quando nos mantemos fiéis a nós mesmos”
(Friedrich Nietzsche)

RESUMO

OLIVEIRA, Taylor Santos. **Desenvolvimento e Verificação do Módulo de Análise de Sentimentos – Nível de Documento Da SentimentALL**. 2019. 76 f. Trabalho de Conclusão de Curso (Graduação) – Curso de Ciência da Computação, Centro Universitário Luterano de Palmas, Palmas/TO, 2019/1.

Dado o rápido crescimento das mídias sociais, os sentimentos expressos pelos usuários a respeito dos mais variados assuntos tornaram-se mais visíveis. As mídias sociais transformaram-se rapidamente em uma verdadeira plataforma de informação e comunicação instantânea, registrando publicamente pensamentos, opiniões e emoções. O *TripAdvisor*® é um exemplo de site de viagens que fornece informações e conteúdos relacionados ao turismo e disponibiliza espaço para que seus usuários avaliem os produtos comprados ou serviços contratados por meio de uma escala de valor ou a partir de comentários. A partir disso, tem-se uma série de pesquisas que buscam formas automatizadas capazes de obter opiniões e atribuir polaridades (positiva, negativa ou neutra) a esses conteúdos. A Análise de Sentimentos é a parte da computação responsável por lidar com esse desafio usando processamento de linguagem natural. Com base no contexto de Análise de Sentimentos, este trabalho abordou o desenvolvimento de um módulo de Análise de Sentimentos baseado no nível do documento a partir de uma adaptação do modelo cascata para analisar avaliações feitas por usuários do *TripAdvisor*®. A aplicação das métricas *precision, recall* e *F-measure*, além da análise de correlação entre o resultado do módulo de Análise de Sentimentos no nível do documento e o valor expresso pelo usuário via escala *Likert* na avaliação realizada no *TripAdvisor*® mostrou que o módulo de AS - Nível do Documento apresentou bons resultados.

Palavras-chave: Análise de sentimento, nível de documento, estatística, polaridade geral

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AS – Análise de Sentimentos

IDE – *Integrated Development Environment*

SQL – *Structured Query Language*

SGBD – Sistema de Gerenciamento de Banco de Dados

NLP – *Natural Language Processing*

PoS – *Part-of-Speech*

EM - *Expectation-Maximization (EM)*

SVM - *Support Vector Machines*

CSV - Comma-separated values

PMI - *Pointwise Mutual Information*

CEULP – Centro Universitário Luterano de Palmas

SUMÁRIO

1 INTRODUÇÃO

A Análise de Sentimentos (AS), ou Mineração de Opinião, é a área de estudo que analisa as opiniões, sentimentos, avaliações, apreciações, atitudes e emoções das pessoas em relação a entidades como produtos, serviços, organizações, indivíduos, questões, eventos, tópicos e todos os seus atributos relacionados (LIU, 2012). Segundo Benevenuto, Ribeiro e Araújo (2015), o principal objetivo da AS é definir técnicas automáticas capazes de extrair informações subjetivas de textos em linguagem natural, como opiniões e sentimentos, a fim de criar conhecimento estruturado que possa ser utilizado por um sistema de apoio ou tomador de decisão.

A Análise de Sentimentos pode ser realizada em diferentes granularidades: em documento, sentença e aspectos. Dentre as granularidades, a AS no nível de documento é a responsável por apresentar uma visão geral sobre os sentimentos expressos no documento como um todo. Por exemplo, ao realizar a avaliação de um produto ou serviço utilizado, o usuário pode apresentar uma opinião para cada característica do produto ou serviço, podendo indicar um sentimento diferente sobre cada uma delas. Os sentimentos expressos sobre cada ponto do produto ou serviço em uma única avaliação podem ser interpretados para dar origem a um sentimento global. Ou seja, as opiniões expressas em cada documento são assumidas como uma única entidade, de forma que o sentimento global pode ser classificado como positivo ou negativo.

O módulo de Análise de Sentimentos desenvolvido neste trabalho é parte do projeto *SentimentALL* de Brito et al. (2015) do grupo de pesquisa Engenharia Inteligente de Dados do CEULP/ULBRA. O projeto objetiva a utilização de técnicas computacionais de Análise de Sentimentos e de técnicas psicológicas de Análise Comportamental aplicadas no estudo do contexto do Turismo Nacional. O ambiente estudado neste projeto é o site especializado em turismo *TripAdvisor*® (https://www.tripadvisor.com.br/). O *TripAdvisor*® é um site de viagens que fornece informações e opiniões de conteúdos relacionados ao turismo.

A partir da pesquisa de Brito (2015), a *SentimentALL* versão 2 foi implementada em Araújo (2017), com o objetivo de extrair, analisar e classificar aspectos presentes em comentários do site de turismo *TripAdvisor*®. Araújo (2017) utiliza análise de relações sintáticas de sentenças para identificar opiniões e aspectos, esse processo tem como base a ideia de que entidades particulares de uma sentença são dependentes para formar um contexto que tenha significado. Para extrair informações de dependências, Araújo (2017) utilizou a

ferramenta *MaltParser* que, a partir de uma base de dados marcados e técnicas de aprendizado de máquina supervisionado, gera um modelo de classificação que identifica dependências entre as palavras de uma sentença. As palavras opinativas receberam classificação de polaridades utilizando os léxicos *SentiLex-PT*, *LIWC* e *OpLexicon*. O resultado desse módulo de classificação de aspectos é uma lista para cada comentário analisado, em que cada item contém uma opinião presente no comentário, a polaridade da opinião e o aspecto relacionado.

A capacidade de classificar sentimentos em vários níveis é importante, pois aplicações diferentes têm necessidades diferentes (MCDONALD *et al.*, 2007). Por exemplo, um sistema de resumo para análises de produtos pode exigir uma classificação de polaridade no nível da sentença ou aspecto, sendo possível identificar a opinião sobre uma parte específica do produto ou serviço avaliado. Já um sistema que determina quais artigos de uma fonte de notícias on-line são de natureza editorial exigiria uma análise de nível de documento (MCDONALD *et al.*, 2007). Em razão disso, é importante que a *SentimentALL* realize a Análise de Sentimento em diferentes granularidades, visto que o resultado da análise no nível do documento apresenta informações sobre uma ótica diferente em relação ao nível do aspecto (vice-versa). O modelo de dados da *SentimentALL* permite relacionar os aspectos às sentenças e essas ao comentário. A partir disso é possível explorar a Análise de Sentimentos no nível do documento. Dessa forma, ao explorar a Análise de Sentimentos na granularidade do documento para a *SentimentALL* espera-se uma ampliação das potencialidades da ferramenta (BRITO, 2018).

Com a Análise de Sentimentos no nível do aspecto e documento dos comentários e avaliações dos usuários de hotéis, restaurantes e atrações do *TripAdvisor*®, é possível identificar os gostos, necessidades e as expectativas dos clientes. Essas informações podem auxiliar no processo de criação de serviços ou produtos que atendam diretamente um cliente ou grupo específico, além de oferecer auxílio em campanhas de *marketing*.

Por esse motivo, o Módulo de Inferência da Polaridade Geral dos Comentários do *TripAdvisor*® analisados na *SentimentALL* versão 2 foi desenvolvido em Oliveira (2018). Com base nos aspectos e polaridades obtidas como resultado no trabalho desenvolvido em Araújo (2017), Oliveira (2018) desenvolveu um modelo estatístico no qual foi possível inferir uma polaridade geral ao comentário. Oliveira (2018) realiza a definição dos pesos dos aspectos, e em seguida, considerando os pesos dos aspectos e suas polaridades, aplica seu modelo estatístico para inferir uma tendência de polaridade (positiva ou negativa) ao comentário como um todo.

A partir disso, foi necessário avaliar o quão efetivo é o módulo de inferência da polaridade geral para inferir corretamente o status positivo ou negativo a um comentário. Com esse entendimento, foi possível através de um levantamento bibliográfico encontrar uma abordagem intitulada “modela cascata” na metodologia apresentada por Zhang et al (2011) para auxiliar o processo de modificação do módulo de inferência da polaridade geral dos comentários.

A abordagem cascata é utilizada por Zhang et al (2011) para AS - Nível do Documento de artigos escritos na língua chinesa. O modelo cascata consiste na divisão do documento em sentenças. Cada sentença é polarizada como positiva ou negativa. Em seguida, o grau de importância (peso) da sentença para o documento é calculado seguindo os seguintes critérios: 1) a posição da sentença, 2) os termos com pesos na sentença, 3) a semelhança entre a sentença e o título, 4) a ocorrência de palavras-chave e 5) o modo de primeira pessoa. Por fim, com o entendimento da importância de cada sentença, a polaridade do documento é inferida.

Portanto, o objetivo geral deste trabalho é criar e testar o módulo de Análise de Sentimentos baseada no nível de documento para ferramenta *SentimentALL* com base no módulo de inferência da polaridade geral dos comentários. Neste sentido objetiva-se especificamente:

- utilizar uma adaptação da abordagem de cascata de AS - nível de documento no processo de definição da polaridade geral de comentários;
- aplicar métricas de avaliação para verificar a qualidade do resultado do módulo de AS - nível de documento;
- analisar a correlação entre a avaliação geral dos comentários via escala *Likert* do *TripAdvisor*® (dados presentes na base de dados da *SentimentALL*) e a polaridade geral detectada pelo módulo AS - nível de documento.

Nesse cenário, este trabalho teve como principal contribuição a verificação da qualidade do resultado do módulo de AS - nível de documento da *SentimentALL*, além de adaptar parte do modelo cascata para a inferência da polaridade geral dos comentários, modificando o peso do aspecto (para maior) de acordo com o entendimento da sua para o comentário.

Este trabalho está estruturado da seguinte forma. O Capítulo 2 apresenta no referencial teórico um estudo sobre os assuntos abordados no trabalho: (2.1) análise de sentimentos ou mineração de opinião, (2.2) análise de sentimento nível do documento, (2.3) escala de *Likert*, (2.4) matriz de confusão e (2.5) correlação. No capítulo 3 é apresentada a metodologia e os materiais utilizados no desenvolvimento do trabalho. O capítulo 4 apresenta os resultados

obtidos, alguns dados são apresentados e discutidos. Por fim as considerações finais são apresentadas no capítulo 5, e as referências bibliográficas utilizadas, no capítulo 6.

2 REFERENCIAL TEÓRICO

Ferramentas que identificam sentimentos em opiniões (por exemplo, opiniões positivas ou negativas sobre produtos ou serviços) na internet são de grande importância no meio corporativo (KAUER, 2016). Liu (2010) diz que essas informações textuais amplamente disponíveis na internet em linguagem natural podem ser categorizadas em dois tipos: opiniões e fatos. Ainda segundo o autor, fatos relatam acontecimentos passados ou verdades, enquanto as opiniões referem-se a pontos de vista subjetivos, expressos sobre as mais diversas coisas, fatos e pessoas.

A partir disso, tem-se uma série de pesquisas que buscam formas automatizadas capazes de obter opiniões e atribuir polaridades (positiva ou negativa) a esses conteúdos. (PANG, LEE, 2004; MCDONALD et al., 2007; BIBI 2017). A Análise de Sentimentos é a parte da computação responsável por lidar com esse desafio utilizando processamento de linguagem natural.

Desde os anos 2000, a Análise de Sentimentos tem crescido e se tornado uma das áreas de pesquisa mais ativas no campo de processamento de linguagem natural (ZHANG, WANG, LIU 2018). A disponibilidade de bases de dados para extração, treino e análise de informações contribuiu para o aumento dessas pesquisas. Este capítulo provê inicialmente uma visão geral dos conceitos e técnicas utilizadas na Análise de Sentimentos, em seguida são apresentados conceitos e trabalhos relacionados a Análise de Sentimentos no nível do documento.

2.1 ANÁLISE DE SENTIMENTOS OU MINERAÇÃO DE OPINIÃO

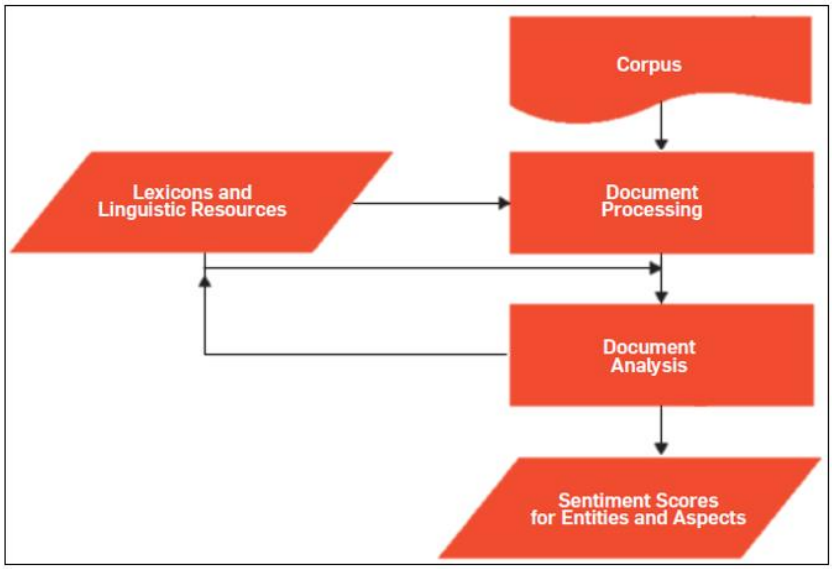
“Análise de sentimentos é a prática de aplicar processamento de linguagem natural e técnicas de análise de texto para identificar e extrair informações subjetivas de textos escritos em língua natural” (HUSSEIM, 2016, p. 1). Segundo Liu (2012), a Análise de Sentimentos (AS) ou mineração de opinião é um campo de estudo que analisa opiniões, sentimentos, avaliações e emoções sobre entidades como produtos, serviços e empresas. Uma opinião pode representar um sentimento, avaliação, atitude e emoção.

A Análise de Sentimentos possui duas tarefas básicas: o reconhecimento de emoção, que tem como foco a identificação de um conjunto de emoções contidas em um *corpus* e a detecção de polaridade, que é responsável pela classificação (geralmente binária) de um aspecto ou *corpus*, produzindo resultados como “positivo” ou “negativo” (CAMBRIA *et al*., 2017).

Ainda Liu (2012) reporta a Análise de Sentimentos como um conjunto de termos, sendo eles: Objeto, Componente, Opinião, Polaridade e Tempo. O objeto é o alvo de análise, pode referir-se a um produto, serviço, pessoa ou uma entidade. O componente refere-se às características do objeto, ou seja, uma opinião pode ser dada sobre um determinado produto, mas ao mesmo tempo sobre uma característica do mesmo. A opinião é a expressão, atitude ou emoção emitida por alguma entidade, a polaridade que determina se a opinião é positiva, negativa ou neutra (BRITO, 2016) e, por último, o tempo, é um momento específico no tempo em que a opinião foi expressa.

Feldman (2013) define a arquitetura geral de um sistema de Análise de Sentimentos como apresentado na figura a seguir.

Figura 1: Arquitetura geral de um sistema genérico de Análise de Sentimentos



Fonte: Feldman (2013)

A entrada do sistema é um corpus, ou conjunto de documentos a serem analisados. Sobre estes documentos, o módulo de processamento do documento utiliza um conjunto de recursos linguísticos para efetuar tarefas tais como tokenização, lematização, ou marcações PoS. Após o processamento do documento, o módulo "Análise do Documento" utiliza os dados processados, juntamente com lexicons de sentimentos, para etiquetar os documentos com as polaridades das opiniões detectadas. Esta análise pode ser efetuada no documento como um todo, para cada sentença ou para cada aspecto observado. Os níveis de análise serão abordados na seção a seguir. Para apresentar os resultados globais da análise ao usuário, o sistema possui

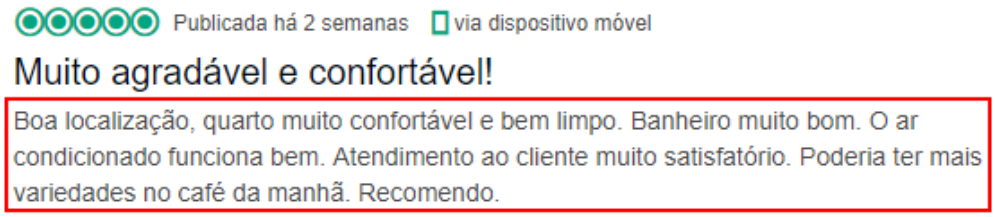
um módulo de pontuação de sentimentos, que é responsável por contabilizar as polaridades anotadas na etapa anterior.

2.1.1 GRANULARIDADE DA ANÁLISE DE SENTIMENTO

A Análise de Sentimentos pode ser realizada em três diferentes granularidades (níveis): Nível de documento, Nível de sentença e Nível de aspecto.

O Nível de documento consiste em classificar a revisão obtida de um documento e indicar o sentimento que todo o documento expressa, podendo ser um valor de sentimento positivo ou um valor de sentimento negativo (LIU, 2012). Ou seja, quando um documento expressa uma opinião geral (positiva ou negativa) sobre uma entidade (produto, serviço, evento etc) (KAUER, 2016). Um contexto interessante em que a análise em nível de documento pode ser utilizada é em *reviews* de produtos ou filmes (LIU, 2010), em vista que esse tipo de avaliação do usuário tende a assumir um sentimento único. Isto é, geralmente nesse tipo de análise o usuário gosta do todo ou não. A Figura 2 apresenta um comentário de um usuário do *TripAdvisor®*. Nesse contexto, o comentário é entendido como um documento, e pode conter vários aspectos e sentimentos expressos.

Figura 2: Comentário (documento)



Fonte: www.TripAdvisor®.com.br (2018)

A Tabela 1 apresenta o comentário da Figura 2 e sua polaridade geral inferida a partir da análise manual realizada pelo autor.

Tabela 1: Polaridades inferidas com análise manual

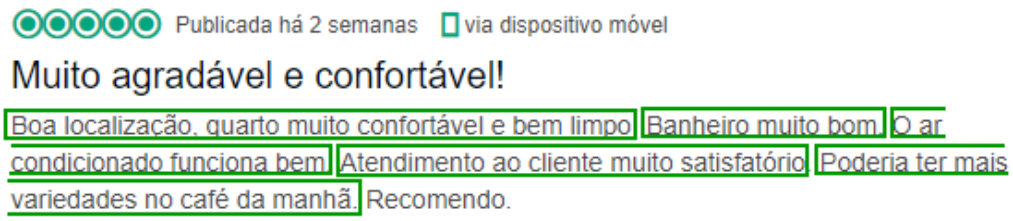
Comentários de usuários do *TripAdvisor®*	
O hotel tem as instalações novas, porém ainda marca o antigo dono do hotel, não remetendo ao Ibis, o atendimento ainda parece que estão aprendendo, e poucos funcionários para atender as demandas.	NEGATIVO

A classificação da polaridade do comentário apresentado na Figura 2 foi realizada por um ser humano a partir da leitura e interpretação do texto. Isso é, uma análise manual. Interpretar sentimentos expressos em um texto não é uma tarefa complexa para um ser humano,

considerando que possuímos capacidade racional e intelectual que pode ser trabalhada para realizar essa atividade. No entanto, tratar isso de forma computacional é muito mais complexo, sendo necessário fazer o uso de Processamento de Linguagem Natural para Análise de Sentimentos no nível do documento, de tal forma que seja possível simular a capacidade do ser humano de interpretar os vários sentimentos que podem estar contidos no documento e classificá-lo como positivo ou negativo. Algumas técnicas que tratam da Análise de Sentimentos na granularidade do documento de forma computacional presentes na literatura serão apresentadas nas seções posteriores.

O **Nível de sentença** é quando cada (frase, cláusula) expressa uma opinião a respeito de um ou vários aspectos (preço, qualidade etc) de uma entidade (KAUER, 2016). Ou seja, implica que a Análise Sentimento é realizada sobre frases ou sentenças, em vez de executada sobre todo o documento. Nesta granularidade cada frase mostra um nível de sentimento positivo ou negativo. A Figura 3 apresenta um comentário (documento) dividido em sentenças.

Figura 3: Comentário (documento) dividido em sentenças



Fonte: www.TripAdvisor®.com.br (2018)

A Tabela 2 apresenta as sentenças do comentário da Figura 3 bem como suas polaridades deduzidas a partir de uma análise manual.

Tabela 2: Comentário dividido em sentenças e polaridades inferidas com análise manual

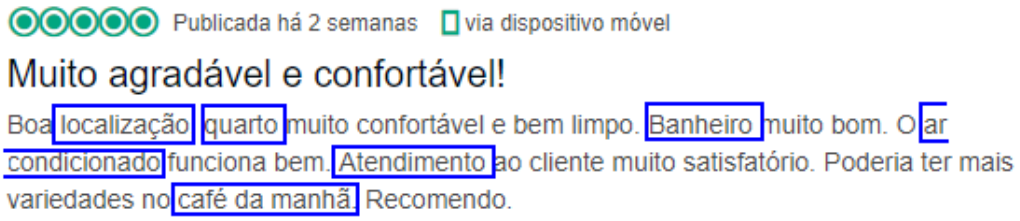
Comentário do usuário do *TripAdvisor*® dividido em sentenças		
Sentença - 1	Boa localização, quarto muito confortável e bem limpo	POSITIVO
Sentença - 2	Banheiro muito bom	POSITIVO
Sentença - 3	O ar condicionado funciona bem	POSITIVO
Sentença - 4	Atendimento ao cliente muito satisfatório	POSITIVO
Sentença - 5	Poderia ter mais variedades no café da manhã	NEGATIVO

A análise manual efetuada sobre as sentenças do comentário da Figura 3 apresentadas na Tabela 2, foram realizadas pelo próprio autor. Foi inferida uma polaridade (positiva ou negativa) as sentenças a partir da interpretação do sentimento expresso pelo usuário. A Tabela

2 tipifica o resultado esperado de um sistema que realiza a Análise de Sentimentos no nível da sentença.

No **Nível de aspecto**, uma sentença pode ser julgada por várias entidades e pode conter múltiplos sentimentos associados a ela (BENEVENUTO, RIBEIRO, ARAÚJO, 2015). Por exemplo, a sentença "Esse hotel, apesar de possuir um ótimo café da manhã, tem uma localização péssima!" possui duas diferentes polaridades, uma associada a "café da manhã" e outra a "localização" para o mesmo hotel. Enquanto para "café da manhã" existe uma polaridade positiva, "localização" possui uma polaridade negativa. Um documento pode conter opiniões sobre diversos aspectos de diversas entidades. "Nessa granularidade é obtido o máximo de detalhes sobre o que o usuário gosta e o que não gosta." (BENEVENUTO, RIBEIRO, ARAÚJO, 2015, p. 5). A Figura 4 apresenta alguns exemplos de aspectos em um comentário de uma avaliação de um hotel no *TripAdvisor*®.

Figura 4: Aspectos identificados no comentário (documento)



Fonte: www.TripAdvisor®.com.br (2018)

A Tabela 3 apresenta os aspectos do comentário da Figura 4 e suas polaridades obtidos com uma análise manual realizada pelo próprio autor.

Tabela 3: Aspectos no comentário e polaridades inferidas com uma análise manual

Aspectos identificados no comentário do usuário do *TripAdvisor*® e suas polaridades		
Aspecto - 1	Localização	POSITIVO
Aspecto - 2	Quarto	POSITIVO
Aspecto - 3	Banheiro	POSITIVO
Aspecto - 4	Ar Condicionado	POSITIVO
Aspecto - 5	Atendimento	POSITIVO
Aspecto - 6	Café da manhã	NEGATIVO

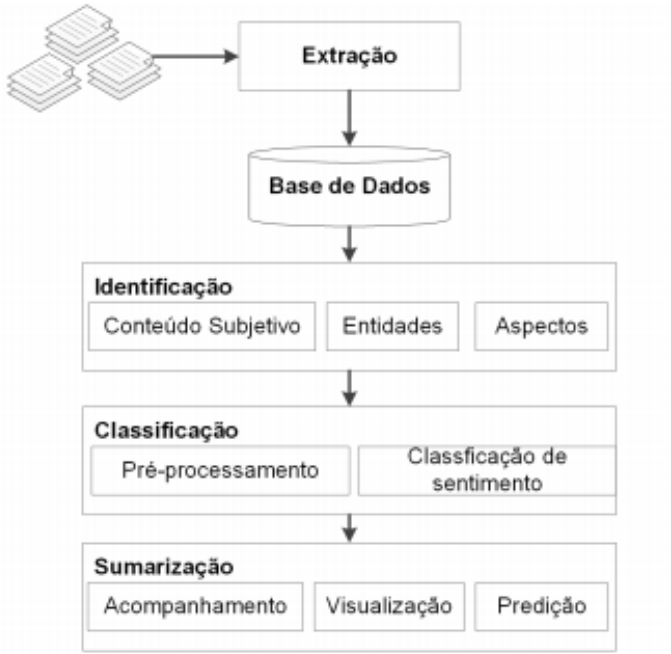
Na Tabela 3, cada aspecto contido no comentário foi polarizado como positivo ou negativo. Esse é um exemplo do resultado esperado de um sistema que realiza uma Análise de

Sentimento no nível do aspecto. Com o intuito de exemplificar, as polaridades dos aspectos do comentário da Figura 4 apresentadas na Tabela 3 foram inferidas a partir de uma análise manual realizada pelo autor.

2.1.2 ETAPAS DA ANÁLISE

Becker e Tumitan (2013) definem três etapas para a mineração de opiniões: Identificação, Classificação de Polaridade e Sumarização. Este processo é esboçado na Figura a seguir.

Figura 5: Etapas da Mineração de Opinião



Fonte: Becker e Tumitan (2013)

2.1.2.1 IDENTIFICAÇÃO

"A etapa de identificação consiste em encontrar os tópicos existentes (e possivelmente seus aspectos), e associá-los com o respectivo conteúdo subjetivo" (BECKER, TUMITAN, 2013, p. 6). A forma de identificar as entidades, aspectos e sentimentos são dependentes da granularidade escolhida para análise. Esta tarefa também pode envolver o discernimento entre documentos ou sentenças com ou sem opinião, visando melhorar os resultados da próxima etapa. Isto é bastante comum quando o nível de análise é de granularidade menor.

2.1.2.2 CLASSIFICAÇÃO DA POLARIDADE

Assim como na fase de identificação, a forma de classificação da polaridade, bem como os resultados obtidos, é relativa à granularidade (nível da Análise) e técnicas utilizadas. Para a classificação da polaridade, diferentes abordagens são propostas na literatura, bem como a abordagem baseada em dicionário, também denominada léxica ou linguística (SILVA, LIMA E BARROS, 2012), abordagem baseada em Aprendizado de Máquina (KALAIVANI, SHUNMUGANATHAN, 2014) e abordagens Estatísticas e Semânticas (BLEI et al., 2003).

Becker e Tumitan (2013) afirmam que, independente da abordagem empregada, a classificação da polaridade não é um problema simples. Os autores ainda apontam os principais desafios:

- a opinião pode depender do observador. Por exemplo, a opinião representada na sentença “As ações da Petrobrás subiram” é positiva para quem detém este tipo de ação, mas pode ser péssima para quem deixou de investir nelas;
- muitos domínios são caracterizados pelo uso frequente de ironias e sarcasmo, onde o sentido implícito é exatamente oposto ao sentimento expresso explicitamente. Outros domínios (e.g. debates políticos, críticas culturais) estabelecem uma opinião positiva por oposição a uma argumentação negativa (ou vice-versa).
- duplo sentido ou ambiguidade de uma palavra no contexto da frase, por exemplo: “João sentou na cadeira e quebrou seu braço”. O braço quebrado é de João ou da cadeira? Casos assim dificultam a inferência da polaridade ao objeto correto.

Há também opiniões que não são facilmente inferidas como positivas ou negativas, e quando processadas o algoritmo não consegue identificar e rotulá-las (OLIVEIRA, 2018). Como por exemplo a frase "Wikinotícias é uma fonte de notícias genérica", o genérica faz da fonte de notícias que é a wikinotícias algo bom ou ruim? Quando não há informação suficiente para definir uma polaridade, considera-se que a opinião é neutra.

2.1.2.3 SUMARIZAÇÃO

Liu (2012) explica que para identificar a opinião média ou prevalecente, a opinião expressa por um pequeno grupo de pessoas não é suficiente, sendo necessário analisar uma grande quantidade de opiniões. Segundo o autor, medidas quantitativas devem fornecer o percentual de pessoas que opinaram positivamente ou negativamente sobre quais aspectos de quais objetos, além dos aspectos mais comentados positiva ou negativamente. Ou seja, é

necessária a criação de métricas e sumários que quantifiquem a diversidade de opiniões encontradas a respeito de um mesmo alvo, podendo ajudar um consumidor a identificar seus respectivos pontos fortes e fracos (BECKER, TUMITAN, 2013). Este é o objetivo desta etapa, onde são criadas métricas que representam o sentimento geral levando em consideração a experiência prévia de outras pessoas, expressas em suas opiniões, as quais podem ser visualizadas ou servir de entrada para outras aplicações.

2.1.3 GERAÇÃO DE LÉXICOS DE SENTIMENTOS

De acordo com Feldman (2013), existem três abordagens para gerar os léxicos de sentimentos: a abordagem manual, a abordagem baseada em dicionário e abordagem baseada em corpus.

- abordagem manual: essa abordagem exige muito esforço e tempo, por esse motivo não é utilizada isoladamente. No entanto, ela é utilizada junto a outras abordagens automatizadas a fim de servir de base para a realização de testes e correções;
- Segundo Liu (2010), a abordagem baseada em dicionário é um processo automatizado para extrair palavras de dicionários online. No primeiro momento é construída uma lista de palavras rotuladas (opinativas e polarizadas), que são chamadas de sementes. Com base nessa lista, os sinônimos e os antônimos das palavras são pesquisados e coletadas em dicionários, recebendo a mesma polaridade da semente.
- A abordagem baseada em *corpus* é um método automatizado que depende de padrões sintáticos e de uma lista de sementes para encontrar palavras opinativas em um *corpus* (LIU, 2012). Silva, Lima e Barros (2012) apresentam o processo para desenvolver uma base léxica.

2.2 ANÁLISE DE SENTIMENTO NÍVEL DO DOCUMENTO

Segundo Bibi (2017) e Liu (2012), a Análise de Sentimentos no nível do documento consiste em classificar a revisão obtida de um documento e indicar o sentimento que todo o documento expressa, podendo ser um valor de sentimento positivo ou um valor de sentimento negativo. Por exemplo, as opiniões que são expressas por um usuário na avaliação de um produto ou serviço são assumidas como uma única entidade, dessa forma, o comentário da avaliação ganha uma classificação geral (positiva ou negativa).

Diferentes abordagens podem ser utilizadas para classificar um documento, por exemplo: aprendizado de máquina e abordagens estatísticas, que incluem *Nave Bayes*, *Support*

Vector Machines (SVM) (BIBI, 2017), além do Modelo Cascata (PANG, LEE, 2004; MCDONALD et al., 2007).

2.2.1 ABORDAGENS E MODELOS ESTATÍSTICOS

McDonald et al (2007) e Pang e Lee (2004) apresentam em seus trabalhos uma abordagem intitulada Modelo Cascata. Albornoz et al. (2017) e Lopes, Oliveira e Vieira (2011) ressaltam que o primeiro passo para atribuir uma classificação geral de polaridade para um documento é extrair os aspectos e classificá-los com informações de polaridade.

Os autores utilizam um modelo estatístico para atribuir uma polaridade geral ao documento. O Modelo Cascata tem como premissa a identificação e classificação da polaridade das sentenças que compõem o documento, para que então possa ser inferido uma polaridade ao documento como um todo.

McDonald et al (2007) e Pang e Lee (2004) utilizam as abordagens SVM, MIRA e *Naive Bayes* para classificar as sentenças dos documentos a serem analisados. McDonald et al (2007) descreve o Modelo Cascata da seguinte forma: um classificador a nível de sentença é executado primeiro, em seguida, os resultado são inseridos em um classificador de nível de documento.

2.3 ESCALA DE LIKERT

Segundo Júnior e Costa (2014, p. 5) "a escala de verificação de Likert consiste em tomar um modelo e desenvolver um conjunto de afirmações relacionadas à sua definição, para as quais os respondentes emitirão seu grau de concordância". Esse modelo foi proposto por Rensis Likert (1932) para mensurar atitudes do contexto das ciências comportamentais. A tabela a seguir mostra um exemplo desta escala para medição de satisfação com um serviço, em 5 pontos.

Tabela 4: Escala de Likert

Estou satisfeito com o serviço recebido				
Discordo totalmente	Discordo parcialmente	Não concordo nem discordo	Concordo parcialmente	Concordo totalmente
1	2	3	4	5

Fonte: Júnior e Costa (2014, p. 5)

Nesta escala os respondentes se posicionam de acordo com uma medida de concordância atribuída ao item, e, de acordo com esta afirmação, se infere a medida do modelo. A escala original tinha a proposta de ser aplicada com cinco pontos, variando de discordância total até a concordância total. Entretanto, atualmente existem modelos chamados do tipo Likert com variações na pontuação, a critério do pesquisador.

Outro exemplo de utilização da escala de Likert é apresentada na Figura a seguir. O uso da escala de Likert é utilizada como parte do formulário de avaliação disponibilizado pelo *TripAdvisor*® para os usuários realizarem avaliações dos hotéis, restaurantes ou atrações das quais utilizaram o serviço.

Figura 6: Utilização da escala de Likert na avaliação no TripAdvisor®

O usuário avaliador pode pontuar a escala de 1 a 5. Sendo que para a nota 1 é atribuído o resultado horrível, para nota 2 o resultado ruim, nota 3 na escala representa um serviço tido como razoável, nota 4 representa um serviço muito bom e nota 5 é entendido como um serviço excelente.

2.4 MATRIZ DE CONFUSÃO

A matriz de confusão fornece a base para descrever a precisão da classificação e caracterizar os erros, ajudando a refinar a classificação. De uma matriz de confusão podem ser derivadas várias medidas de precisão da classificação, sendo a exatidão global uma das mais conhecidas (FOODY, 2002).

“A matriz de confusão é formada por um arranjo quadrado de números dispostos em linhas e colunas que expressam o número de unidades de amostras de uma categoria particular relativa” (FIGUEIREDO, VIEIRA, 2007, p. 2) – inferida por um classificador (ou regra de decisão), comparado com a categoria atual verificada no campo (CONGALTON, 1991).

Para exemplificar, será apresentado um exemplo de utilização da Matriz de Confusão.

Tabela 5: Matriz de Confusão

		Verdade de Campo			Total	Inclusão	Pixels bem classificados [%]
		A	B	C			
Classes do Mapa Temático	A	35	2	2	39	10,2	89,8
	B	10	37	3	50	26,0	74,0
	C	5	1	41	47	12,8	87,2
Total pixels de campo		50	40	46	136		
Omissão [%]		30,0	7,5	10,9			Exatidão Global [%] 83,1

Fonte: Adaptada de Richards, J. A. (1986, p. 272)

A Tabela 5 apresenta uma matriz de confusão com três classes (A, B e C) A matriz de confusão restringe-se às linhas e colunas referentes às classes A B e C. A diagonal principal apresenta os elementos (número de pixels) que foram classificados corretamente em cada classe (MARTINS, 2012). Exemplo, para a classe A, no modelo do mapa temático foram classificados 35 pixels, na classe B 37 pixels e na classe C 41 pixels respectivamente classificados de forma correta.

No entanto, para 5 pixels da classe C no modelo do mapa temático, em contraste a verdade de campo, demonstrou que na verdade eles faziam parte da classe A (MARTINS, 2012). Da mesma forma 2 pixels classificados como fazendo parte da classe A no modelo do mapa temático, eram na verdade parte da classe C.

Com isso, Martins (2012) coloca que para todos os pixels pertencentes a classe B no modelo do mapa temático, 37 de fato pertenciam a classe B em comparação a verdade de campo, enquanto o restante – 13 – foram classificados incorretamente. “Este erro de classificação é denominado **erro de inclusão (commission)**, pois se está incluindo pixels em uma classe quando na verdade eles pertencem a outra(s)” (MARTINS, 2012, p. 25). Ao realizar a análise da matriz de confusão, Martins (2012) coloca que na classe C 41 pixels foram bem classificados, porém, existem 2 pixels pertencentes a classe C que foram classificados como

sendo da classe A e outros 3 que na verdade faz parte da classe B. “Este erro agora é o **erro de omissão (omition)**, pois nos dois casos está-se a omitir pixels da classe correta atribuindo-os a outra(s) classe(s)” (MARTINS, 2012, p. 25).

A última coluna apresenta a exatidão de cada uma das classes, isto é, o percentual de pixels do modelo do mapa temático que foram classificados corretamente. Ao final da última coluna, é apresentado a **exatidão global (accurary)**. “Neste caso, tínhamos no total 136 pixels, sendo que no mapa temático 113 foram bem classificados, o que perfaz um percentual de 83,1% do total, que foram bem classificados” (MARTINS, 2012, p. 25).

O coeficiente *Kappa* (κ) é uma métrica que pode ser obtida como um subproduto da matriz de confusão, que auxilia na avaliação do classificador. “Uma das vantagens alegadas para uso do *Kappa* (κ) é de que ele incorpora a informação dos pixels mal classificados, e não apenas dos bem classificados como a exatidão global” (OLIVEIRA, 2003, p. 20). Ainda Oliveira (2003, p. 20) apresenta a equação que fornece o valor de *Kappa* (κ), que é dada por:

$$\hat{k} = \frac{N \sum_{i=1}^{\gamma} x_{ii} - \sum_{i=1}^{\gamma} x_{i+} x_{+i}}{N^2 - \sum_{i=1}^{\gamma} x_{i+} x_{+i}}$$

onde:

- Σ representa o somatório em cada linha e coluna;
- γ é o número de linhas e de colunas;
- *N* é o número total de pontos, o somatório de toda a matriz.

Outras métricas que podem ser exploradas a partir da matriz de confusão são *precision* (precisão), *recall* (revocação) e *f1-score* (medida F). Com as métricas *precision* (precisão), *recall* (revocação) e *f1-score* (medida F) temos uma visão clara dos resultados e nos permite entender melhor como o modelo está funcionando.

Para entender o que é a Precisão, Revocação e F1-Score, primeiro temos que saber sobre a terminologia para classificação. São elas:

- ***True positive* (TP):** significa que o modelo classificou de forma correta os valores positivos que realmente são positivos considerando a verdade de campo.
- ***True negative* (TN):** significa que o modelo classificou de forma correta os valores negativos que realmente são negativos considerando a verdade de campo.
- ***False positive* (FP):** diz respeito aos elementos que o modelo como positivo, porém, na verdade de campo eles são considerados negativos.

- ***False negative*** **(FN):** diz respeito aos elementos que o modelo classificou como negativo, porém, na verdade de campo eles são considerados positivos.

A precisão (*precision*) é calculada da seguinte forma:

$$precision = \frac{true\ positive}{true\ positive + false\ positive}$$

Significa a proporção de positivos classificados corretamente, ou seja, dos classificados como positivos quantos são positivos realmente.

A revocação (*recall*) é calculada da seguinte forma:

$$recall = \frac{true\ positive}{true\ positive + false\ negative}$$

Significa a proporção de positivos identificados corretamente, ou seja, o quão bom o modelo é para detectar positivos

Já a *F1-score* é calculada da seguinte forma:

$$F - Measure = 2 * \frac{precision * recall}{precision + recall}$$

Essa é a média harmônica entre *precision* e *recall*. Com essa informação podemos dizer a performance do classificador com apenas um indicador.

Como essa medida é uma média, ela apresenta uma visão mais exata da eficiência do classificador do que apenas a *precision* ou a *recall*.

2.5 CORRELAÇÃO

A correlação é um dos métodos mais utilizados dentro da estatística para medir o grau de associação entre variáveis, sendo a correlação de *pearson* ou momento produto umas das mais precisas, como mostra Lira (2004).

Segundo o dicionário Aurélio, o termo “correlação” significa semelhança; relação de correspondência entre dois seres, duas coisas, duas ideias que se relacionam entre si, e indica até que ponto os valores de uma variável estão relacionados com os de outra. Na análise de correlação procura-se representar o grau de relacionamento entre as variáveis em um único número

Muitos exemplos podem ser dados de variáveis que apresentam certo tipo de relacionamento: a) grau de escolaridade e nível de renda; b) notas de português e notas de redação; c) idade e resistência física; d) produtividade e quantidade de fertilizantes utilizada; e) ordem de classificação em um concurso e sucesso profissional.

O interessante de se conhecer melhor o relacionamento entre variáveis, como casos citados anteriormente, conduz naturalmente à análise de correlação. O resultado é uma medida do grau de correlação, denominada “coeficiente de correlação”.

A principal utilidade da medida de correlação é que se pode dizer o que se espera para uma variável com base no conhecimento de outra. Pode-se inferir uma com base na outra. Contudo, chama-se a atenção para o fato de que esse processo de inferência não significa que uma variável “causa” a outra. Ou seja, não implica, em hipótese alguma, a existência de relação causal entre as variáveis. Apenas o relacionamento esperado é indicado pela análise de correlação.

2.5.1 CORRELAÇÃO DE PEARSON

Também conhecido como correlação momento produto, procura medir o grau de relacionamento linear entre duas variáveis. A medida usada é o “coeficiente de correlação”.

2.5.1.1 COEFICIENTE DE CORRELAÇÃO DE PEARSON

Coeficiente de correlação ou grau de relacionamento entre as variáveis. A medida desse grau é feita pelo coeficiente de correlação.

Considerando duas variáveis (X e Y) definidas para uma amostra de tamanho n, o coeficiente de correlação de *pearson* é calculado com a seguinte fórmula (LIRA, 2004):

$$p = \frac{\Sigma(x_i - \underline{x})(y_i - \underline{y})}{\sqrt{\left(\Sigma\,(x_i - \underline{x})^2\right)\left(\Sigma\,(y_i - \underline{y})^2\right)}}$$

2.5.1.2 PROPRIEDADES DO COEFICIENTE DE CORRELAÇÃO

Lira (2004) relata que o coeficiente de correlação possui as seguintes propriedades:

1. As unidades de medida da variável não afetam o coeficiente de correlação; é um número expresso que varia entre –1 e +1.
2. O coeficiente de correlação de *pearson* de uma variável e ela mesma é igual ao valor máximo do coeficiente de correlação, ou seja, igual a +1.
3. A ordem das variáveis utilizadas no cálculo da correlação não afeta o resultado do coeficiente de correlação.
4. O coeficiente de correlação não se altera ao somar ou subtrair as variáveis a uma constante.

5. O coeficiente de correlação não se altera ao multiplicar ou dividir as variáveis por uma constante.

2.5.1.3 INTERPRETAÇÃO DO COEFICIENTE DE CORRELAÇÃO

O coeficiente de correlação é expresso no intervalo -1 e +1. Esse intervalo pode de entendido da seguinte forma (LIRA, 2004, p. 41):

- se $0,00 < r < 0,30$, existe fraca correlação linear;
- se $0,30 \leq r < 0,60$, existe moderada correlação linear;
- se $0,60 \leq r < 0,90$, existe forte correlação linear;
- se $0,90 \leq r < 1,00$, existe correlação linear muito forte.

O coeficiente de correlação igual a zero indica ausência de correlação linear entre as variáveis. Pode ocorrer, no entanto, que as variáveis sejam relacionadas, porém não linearmente.

2.4 TRABALHO RELACIONADO

Zhang et al (2011) demonstram quais os procedimentos que foram realizados para fazer uma Análise de Sentimentos no nível do documento de artigos escritos na língua chinesa. O modelo é composto por duas grandes etapas: Análise de Sentimentos da frase e agregação do sentimento do documento. O primeiro passo foi a identificação do sentimento de cada sentença que compunha o documento, e então as sentenças foram consideradas como unidades atômicas para análise semântica.

> A análise semântica trata do significado da sentença, sobre aquilo que é possível entender através de um determinado enunciado. Análise semântica envolve a elaboração de uma representação dos objetos e ações que uma sentença esteja descrevendo, incluindo detalhes fornecidos por adjetivos, advérbios e preposições (COPPIN, 2013, p. 511).

Zhang et al (2011) apresentam que um documento pode ser segmentado em múltiplas sentenças, das quais, a partir de suas polaridades, a polaridade do documento pode ser inferida. Os autores utilizam um dicionário de palavras subjetivas chinesas resumidas por *HowNet* (*Dong* e *Dong*, 2003) e assim relacionam as sentenças ou frases subjetivas. Dessa forma, as

frases ou sentenças que não continham palavras subjetivas foram desconsideradas, visto que segundo os autores elas não mostravam polaridade de sentimento.

No entanto, devido à sutileza da expressão da língua chinesa, simplesmente prever a polaridade de uma sentença com base nas palavras subjetivas que ela contém não é suficiente. Uma análise mais detalhada da estrutura sintática é necessária para determinar a polaridade do sentimento. Para essa tarefa, Zhang et al (2011) utilizaram um analisador de gramática de dependência chinesa de código aberto, denominado HIT-IR LTP, desenvolvido pelo Instituto de Tecnologia de Harbin para converter cada sentença em uma árvore de dependência.

Na árvore de dependência gerada, cada nó representa uma palavra e a árvore é composta de múltiplas relações binárias entre palavras. Cada relação tem uma palavra como pai (ou cabeça) e a outra um como filho (ou modificador). Cada palavra tem um e apenas um pai, enquanto uma palavra pode ter vários filhos, explica Zhang et al (2011).

Para determinar a polaridade de uma palavra subjetiva, os autores levam em conta seus modificadores dependentes na sentença. Cada palavra tem sua polaridade prévia definida com base na palavra subjetiva do dicionário. Ao aparecer em uma frase representada como árvore de dependência, uma palavra pode ou não ter um número de filhos que podem modificar sua polaridade.

Para inferir a polaridade geral do documento, os autores adotaram uma abordagem cascata. Para isso, os autores definiram graus de importância da sentença para o documento baseado em cinco características: 1) a posição da sentença (wp), 2) os termos com pesos na sentença (wt), 3) a semelhança entre a sentença e o título (wh), 4) a ocorrência de palavras-chave (wk) e 5) o modo de primeira pessoa (wf). As cinco características são definidas da seguinte forma:

- **posição da sentença (wp)**: dado um documento D = {s1, s2, ..., sN} sendo um conjunto de frases, onde si é a sentença, a posição de uma sentença em um documento pode indicar sua importância. As sentenças iniciais e finais são frequentemente sentenças temáticas e por isso são consideradas mais importantes que às demais. Assim sendo, são atribuídos pesos mais altos às sentenças nas duas extremidades do documento. O recurso de posição (wp) de uma frase é definido como:

$$wp(s_i) = \frac{1}{min(i, N - i + 1)}$$

onde $1 \leq i \leq n$.

- **os termos com pesos na sentença (wt)**: referem-se às sentenças que contêm os termos mais importantes. A ocorrência de termos em uma sentença pode indicar sua significância em um documento. Ou seja, a importância aumenta proporcionalmente ao número de vezes que uma palavra aparece na frase e no documento.
- **semelhança entre a sentença e o título (wh):** trata da similaridade entre uma frase e o título. Já que o título pode ser considerado um resumo do documento, logo, uma sentença (w) semelhante para ao título (h) deve contribuir mais para o documento.
- **ocorrência de palavras-chave (wk):** mede o número total de palavras-chave que ocorrem na frase. A frequência de palavras-chave na frase também indica a relevância e importância da sentença para o documento. Zhang et al (2011) utilizam a abordagem de extração de palavras-chave de Matsuo e Ishizuka (2004) para identificar palavras-chave no documento.
- **modo de primeira pessoa (wf):** indica se a sentença está em o modo de primeira pessoa. Zhang et al (2011) afirma que frases em modo de primeira pessoa, indicadas por pronomes como eu/e/nós, tendem a ter mesma polaridade do documento inteiro.

Para chegar ao peso da sentença, os cinco parâmetros são somados e determinam o efeito de cada característica no peso total de uma sentença. Para tanto, os autores Zhang et al (2011) rotularam um conjunto de dados de treinamento manualmente para ajustar os valores desses cinco parâmetros. Foram escolhidos 20 documentos aleatoriamente de notícias, blogs e fóruns que abrangem vários domínios, como política, economia e assim por diante. Em seguida, rotuladas manualmente as frases importantes nesses documentos para serem usadas como um conjunto de dados de treinamento para ajustar os cinco parâmetros.

Tendo calculado os pesos e pontuações de polaridade de todos as sentenças, Zhang et al (2011) afirmam que é possível calcular a polaridade do documento a partir da soma ponderada de todas as sentenças:

$$p^d = \sum_{i=1}^{n} w_i \, p_i$$

onde P_i denota a pontuação de polaridade da frase s_i, sendo si um conjunto de frases ou sentenças de um documento.

Os autores comparam o resultado obtido na sua abordagem baseada em regras de relação e árvore de dependência com abordagens de aprendizado de máquina. Para isso, os algoritmos de aprendizado de máquina SVM e *Naive-Bayes* são executados sobre o mesmo

conjunto de dados aplicados no desenvolvimento do modelo baseado em regras de relação e árvore de dependência.

Tabela 6: Precisão das abordagens baseadas em aprendizado no conjunto de dados

SVM	Naive Bayes	Decision tree
80,24%	66,28%	71,45%
81,29%	65,57%	69,92%
79,41%	68,39%	74,50%
80,24%	64,04%	70,27%
82,47%	67,57%	**76,50%**
83,88%	**68,86%**	75,21%
81,29%	67,33%	75,09%

Fonte: Zhang el at. (2011)

A Figura 6 apresenta o resultado obtido por Zhang et al (2011) com o uso do seu modelo que aplica árvore de decisão para atribuir pesos as sentenças e pondera os valores dos pesos das sentenças a fim de atribuir um peso e uma polaridade ao documento obteve um resultado extremamente bom, em comparação ao uso de algoritmos de aprendizagem de máquina executados sob o mesmo conjunto de dados.

A partir do trabalho desenvolvido por Zhang et al (2011), observa-se que parte dos processos apresentados pelos autores podem ser adaptados ao modelo de Oliveira (2018), como por exemplo: considerar várias características apresentadas pelos aspectos para atribuição do seu peso (o modelo proposto por Oliveira (2018) considera apenas a proporção entre o total de análises do aspectos e o total de análises de todos os aspectos), além da possibilidade de atribuir pesos as sentenças que compõem o comentário, e a partir daí realizar a inferência da polaridade geral do comentário de forma estatística.

3 METODOLOGIA

Esta foi uma pesquisa aplicada a fins práticos com o propósito de adaptar a abordagem cascata de Zhang et al (2011) ao módulo de AS - Nível do Documento versão 1 de Oliveira (2018), além de aplicar métricas para verificar a qualidade do resultado alcançado pelo módulo antes e depois da adaptação da abordagem cascata.

Esta seção apresenta os métodos bem como os materiais que foram essenciais para o desenvolvimento deste trabalho.

3.1 MATERIAIS

Para o desenvolvimento do módulo de Análise de Sentimentos no nível do documento versão 2 foram utilizados os seguintes materiais e tecnologias:

- **SQL *Server Enterprise* 2017**: o SQL *Server Enterprise* 2017 foi o sistema de gerenciamento de banco de dados utilizado na manipulação e armazenamento dos dados processados e gerados neste trabalho. Com uma instância do servidor SQL *Server Enterprise* 2017 instalada e configurada na máquina utilizada no desenvolvimento, realizou-se a restauração de um *backup* da base de dados da *SentimentALL*;
- ***Integrated Development Environment* (IDE) DataGrip 2018.3.4**: o DataGrip é um IDE de plataforma cruzada destinado a desenvolvedores que trabalham com bancos de dados SQL. O DataGrip foi utilizado como suporte na manipulação e alterações realizadas no banco de dados;
- **Linguagem *Python* 3.6.4**: *Python* é uma linguagem de programação de alto nível. Foi a linguagem de programação utilizada no desenvolvimento do módulo de AS - Nível do Documento;
- ***Integrated Development Environment* (IDE) PyCharm 2018.3**: a IDE PyCharm foi utilizada como suporte no desenvolvimento, ela permite que a codificação seja mais rápida com o auto completar, possui uma ferramenta de depuração de código, o console interativo do *Python* é integrado à IDE, e possui outras ferramentas que ajudaram o processo de desenvolvimento deste trabalho;
- **Excel:** O *Microsoft Office Excel* é um editor de planilhas. Foi utilizado como suporte a organização dos dados além de ser empregado na realização do cálculo do coeficiente da correlação de *Pearson.*

3.2 BASE DE DADOS

Os dados utilizados neste trabalho são conjuntos de aspectos e comentários que, neste contexto, são atributos de destinos turísticos. Os dados utilizados foram os seguintes:

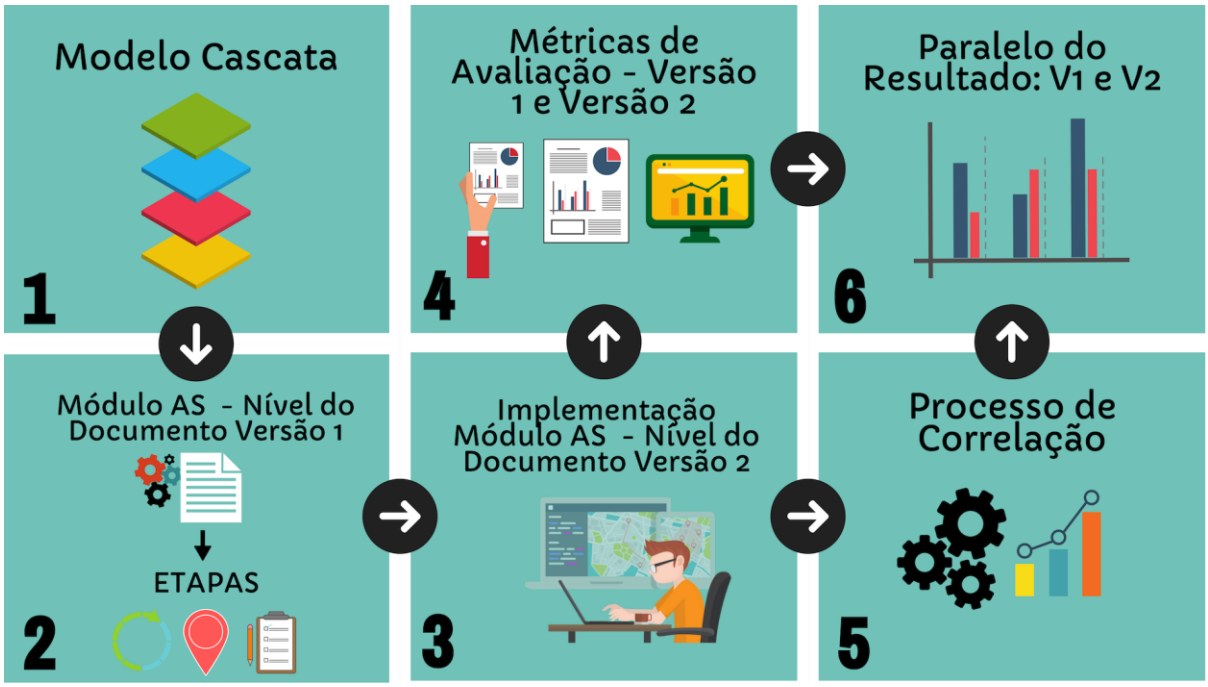
- **Tabela Avaliação**
  - idAvaliação
  - comentário
  - título
  - nota
- **Tabela Análise**
  - idAnálise
  - idSentença
  - aspecto
  - polaridade
- **Tabela Sentença**
  - idSentença
  - idAvaliação
  - seqSentença
  - texto
- **Tabela Aspectos**
  - aspecto
  - total

Os aspectos foram extraídos de avaliações de usuários do site *TripAdvisor*®, a partir da ferramenta *SentimentALL* de Araújo (2017) e que foram avaliados como positivos ou negativos pelo módulo de Análise de Sentimentos desta ferramenta. A *SentimentALL* desenvolvida em Araújo (2017) analisou um total de 4.597.242 comentários, das mais de 6.000.00 avaliações extraídas, onde foram identificados 60.387 aspectos diferentes. Cada aspecto foi classificado no contexto da sua sentença, resultando em 22.000.000 análises. A partir desses dados, foi possível realizar esta pesquisa.

3.3 PROCEDIMENTOS

Para o desenvolvimento e verificação do módulo de análise de sentimentos - nível do documento da *SentimentALL*, foi necessário a realização das seguintes etapas:

Figura 7: Metodologia



1. **Modelo Cascata:** a primeira etapa foi caracterizada pelo estudo e entendimento do modelo cascata utilizado por Zhang et al (2011) para AS - Nível do Documento de artigos escritos na língua chinesa.
2. **Módulo AS - Nível do Documento Versão 1:** a partir do entendimento da forma e aplicação do modelo cascata, a segunda etapa foi marcada pela realização de um paralelo entre o módulo de AS - Nível do Documento Versão 1 de Oliveira (2018) e o modelo cascata. O objetivo era compreender as modificações necessárias a serem realizadas no módulo de AS - Nível do Documento Versão 1 para adaptar o modelo cascata. As modificações foram divididas nas seguintes etapas:
    a. **Análise das etapas:** análise das etapas do modelo cascata possíveis de serem adaptadas ao módulo de AS - Nível do Documento Versão 1 de Oliveira (2018);
    b. **Análise dos dados:** análise dos dados da base de dados da *SentimentALL* a serem utilizados;
    c. **Alterações no banco de dados:** alterações necessárias no banco de dados para comportar a adaptação da abordagem cascata.

3. **Implementação Módulo AS - Nível do Documento Versão 2:** a terceira etapa correspondeu ao desenvolvimento do módulo de AS - Nível do Documento Versão 2. Para isso, foram realizadas modificações no banco de dados para comportar a adaptação do modelo cascata, além de modificar o modelo da definição dos pesos dos aspectos e com base nos critérios utilizados por Zhang et al (2011).
4. **Métricas de Avaliação - Módulo AS - Nível do Documento Versão 1 e Versão 2:** a quarta etapa consistiu na verificação da qualidade do resultado dos módulos de AS - Nível do documento Versão 1 e Versão 2**.** Para isso foram utilizadas as métricas *precision, recall* e *F - measure.* A aplicação das métricas deu-se nas seguintes etapas:
    a. **Seleção dos comentários:** a partir da base de dados da *SentimentALL*, foram selecionadas 150 avaliações de forma aleatória, para que sob as mesmas fosse realizada a análise utilizando as métricas *Precision*, *Recall* e *F-Measure*;
    b. **Organização dos comentários selecionados:** os comentários das avaliações selecionadas foram organizados em planilhas a fim de facilitar a análise dos dados;
    c. **Análise manual:** consistiu na criação de um conjunto de controle, onde a partir dos comentários selecionados foi aplicado o processo de análise de sentimentos de forma manual onde cada comentário foi polarizado como positivo (recebendo o valor 1), negativo (recebendo o valor -1) ou neutro (recebendo o valor 0);
    d. **Criação da matriz de confusão:** para o conjunto de resultados (análise manual – análise equivalente do sistema), foi criada uma matriz de confusão, cujo objetivo é a identificação quantitativa dos comentários positivos que foram classificados como positivos (Verdadeiros-Positivos ou V_P), dos comentários negativos que foram classificados como positivo (Falsos-Positivos ou F_P), dos comentários negativos que foram classificados como negativos (Falsos-Negativos ou F_N) e dos comentários positivos que foram classificados como negativos (Verdadeiros-Negativos ou V_N). Nesse ponto, os comentários classificados como neutros na análise manual ou do sistema foram desconsiderados. Com, das 150 avaliações selecionadas para realizar o processo da matriz de confusão, foram consideradas as 100 primeiras que continham o valor da análise do sistema ou manual diferente de zero.

e. **Aplicação das técnicas "Precision" e "Recall" para análise de desempenho:** utilizando as informações da matriz de confusão, foram aplicadas as fórmulas de "*Precision*", "*Recall*" e *F-Measure* para indicar o índice de desempenho do sistema como um todo.

5. **Processo de Correlação:** na quinta etapa foi realizado o processo de correlação considerando 150 avaliações da base de dados da *SentimentALL*. O editor de planilhas *Excel* foi utilizado para realizar o cálculo do coeficiente de correlação de *Pearson* entre o valor da escala *Likert* expresso pelo usuário na avaliação no *TripAdvisor*® e o resultado obtido pelos módulos de AS - Nível do Documento Versão 1 e Versão 2. As avaliações que continham o valor da análise do sistema igual a zero e o valor da escala *Likert* igual a três foram desconsideradas, pois o valor três está situado no meio da escala *Likert*, e não pode ser considerado totalmente como positivo ou negativo, visto que a escala *Likert* utilizada no *TripAdvisor*® é composta pelos valores um, dois, três quatro e cinco. Além disso, foram consideradas as cem primeiras avaliações restantes após a aplicação dos critérios de neutralidade. Leal, Christhi e Brito (2017) consideraram o total de cem avaliação para a avaliação da qualidade do resultado da SentimentALL. Por esse motivo, também foram consideradas cem avaliações para a análise de correlação e avaliação da qualidade do resultado do módulo de AS – Nível do Documento.
6. **Comparativo do Resultado - Versão 1 e Versão 2:** ao final realizou-se um paralelo dos resultados obtidos entre o módulo desenvolvido no presente trabalho (módulo de AS - Nível do Documento versão 2) com o módulo de AS - Nível do Documento versão 1. Foram comparados os resultados alcançados na análise de correlação, além dos resultados obtidos na avaliação por meio das métricas de *precision*, *recall* e *F - measure* a fim de verificar qual dos módulos obtiveram os resultados mais próximos da análise manual.

4 RESULTADO E DISCUSSÃO

Esta seção apresenta os resultados obtidos no desenvolvimento deste trabalho, bem como as discussões referentes aos resultados. Deste modo, a divisão deste capítulo segue as quatro etapas de divisão do desenvolvimento do trabalho, cada qual apresentando seus resultados específicos e discussões. Na seção 4.1, serão apresentados a arquitetura e o modelo de dados da *SentimentALL*, o modelo de inferência da polaridade geral dos comentários do *TripAdvisor*® analisados na *SentimentALL* de Oliveira (2018) e uma análise do seu desempenho.

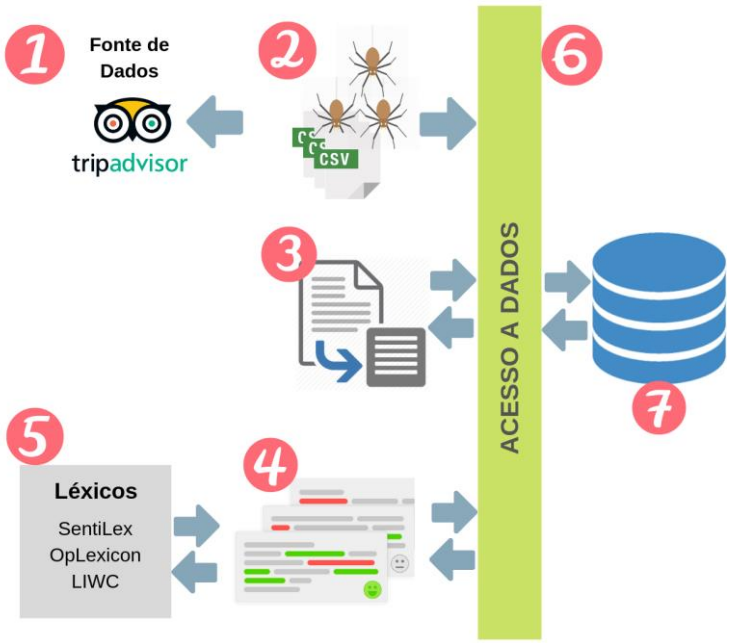
A seção 4.2 é apresenta o aperfeiçoamento do modelo estatístico proposto em Oliveira (2018); a seção 4.3 trata sobre a definição dos pesos dos aspectos; na seção 4.4 é apresentado o processo de inferência da polaridade geral do comentário, na seção 4.5 é discorrido sobre o processo de avaliação do desempenho do módulo.

A seção 4.6 é apresentado o processo de correlação entre o resultado obtido no módulo desenvolvido no presente trabalho com o os dados da escala *Likert* das avaliações presente na base de dados da *SentimentALL*, e por fim, na seção 4.7, é apresentada a comparação entre o módulo desenvolvido no presente trabalho com o módulo de inferência da polaridade geral dos comentários do *TripAdvisor*® desenvolvido em Oliveira (2018).

4.1 Arquitetura *SentimentALL*

O estudo desenvolvido neste trabalho foi realizado com base em análises feitas no módulo de AS - Nível do Documento Versão 1, utilizando a ferramenta *SentimentALL*. Segundo Araújo (2017), essa ferramenta tem como objetivo realizar a mineração de opiniões oriundas de sites da internet escritas em Português do Brasil e aplicar o processo de Análise de Sentimentos. A Figura 8 apresenta a arquitetura da SentimentALL.

Figura 8: Arquitetura SentimentALL



A ferramenta *SentimentALL*, desenvolvida por Araújo (2017) e Brito (2018), utiliza como fonte de dados avaliações sobre destinos no Brasil escritas em Português, extraídas do website *TripAdvisor* (Figura 8-1). O site oferece espaço para que os usuários avaliem acomodações, restaurantes e atrações de uma grande quantidade de destinos no mundo inteiro. Para realizar a coleta desses dados são utilizados *Spiders* (Figura 8-2).

Os *Spiders* ou *Crawlers* são *softwares* capazes de visitar sistematicamente páginas HTML e extrair dessas páginas dados definidos como importantes para o escopo do projeto. No contexto da ferramenta *SentimentALL* são coletados, por exemplo, o texto dos comentários/avaliações sobre destinos turísticos e dados complementares como cidade do Autor, data da publicação da avaliação, pontuação dada pelo usuário por meio da escala *Likert* do site, entre outros. O resultado dessa etapa são documentos de texto no formato CSV. A Figura 8-3 representa a segunda etapa da ferramenta que faz o pré-processamento dos dados e é composta por um conjunto de técnicas da área de processamento de linguagem natural, essas

técnicas são: Normalização de dados, Tokenização, *POS Tagging* e PMI (*Pointwise Mutual Information*). Essa etapa tem como resultado dados estruturados que possuem informações suficientes para que sejam usados na etapa de análise.

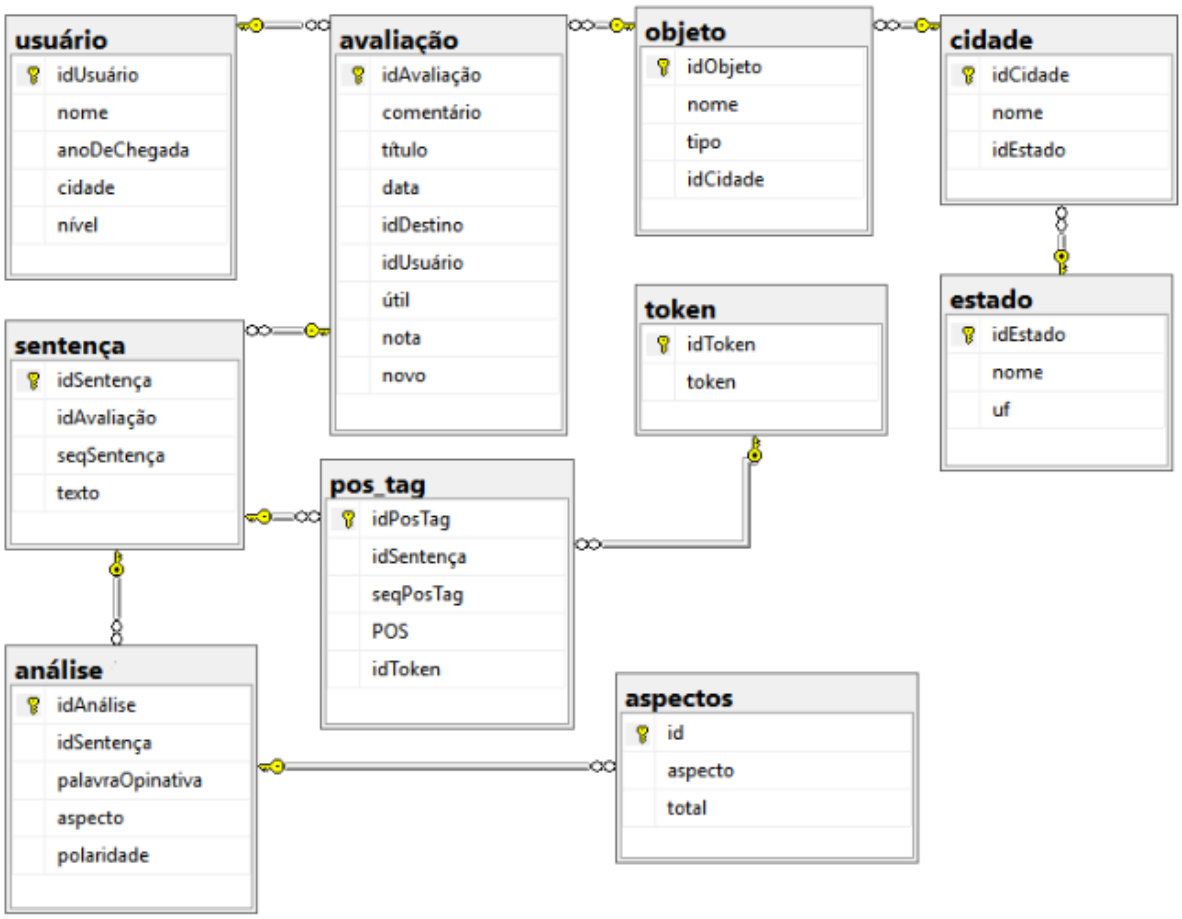
A Figura 8-4 representa a estrutura para o processo de análise de sentimento no nível de aspecto, nessa etapa os comentários que foram processados anteriormente passam por um algoritmo de classificação de Dependências sintáticas. “à noção fundamental de dependências é baseado na ideia de que a estrutura sintática de uma frase consiste de relações binárias assimétricas entre palavras da frase” (NIVRE, 2005, pág. 3, tradução nossa). De forma sucinta, essas dependências definem o relacionamento entre entidades numa frase. Essa estrutura é importante para o processo de identificação da opinião do autor sobre determinado aspecto. Após identificar aspectos e opiniões relacionadas é feita classificação da polaridade dessas opiniões.

A classificação de polaridade é feita com base em léxicos de adjetivos representados na Figura 8-6. Os resultados das etapas do processo de análise de sentimentos no nível de aspecto são enviados para um Banco de Dados (Figura 8-7) através da Camada de Acesso que são funções que auxiliam na obtenção e carga de dados (Figura 8-6).

4.1.1 Banco de dados da *SentimentALL*

A Figura 9 apresenta o modelo relacional completo da *SentimentALL* desenvolvida em Araújo (2017).

Figura 9: Modelo Relacional SentimentALL



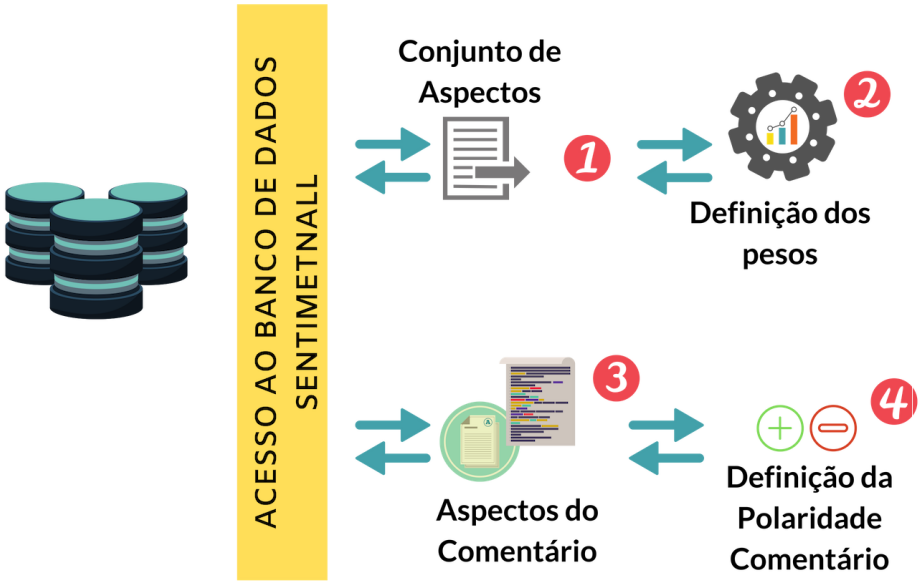
As tabelas **usuario**, **avaliação**, **objeto**, **estado** e **cidade** da Figura 9 foram criadas para armazenar os dados referentes a etapa de coleta dos dados da *SentimentALL*. Na etapa de Pré-Processamento, as tabelas **avaliação**, **sentença**, **pos_tag**, **aspectos** e **token** foram utilizadas. As tabelas **sentença** e **análise** foram manuseadas na etapa de Análise de Sentimentos realizada pela *SentimentALL*.

Para o desenvolvimento do módulo de AS - Nível do Documento desenvolvido neste trabalho, foi necessário o uso dos dados da base de dados da *SentimentALL*. A partir dos dados das tabelas **Aspectos**, **Avaliação, Análise** e **Sentença** foi possível desenvolver uma adaptação do modelo cascata para a AS - Nível do Documento.

4.2 AS - Nível do Documento Versão 1

A estrutura do módulo de AS - Nível do Documento Versão 1 desenvolvida em Oliveira (2018) é apresentada na Figura 10.

Figura 10: Estrutura módulo AS-Nível do Documento Versão 1



O módulo de AS - Nível do Documento Versão 1 tem como entrada um conjunto de aspectos (por exemplo, atendimento, comida, quarto etc.) (Figura 10-1) avaliados como positivos ou negativos extraídos dos comentários de avaliações de usuários do site *TripAdvisor*®, a partir da ferramenta SentimentALL (ARAÚJO, 2017; BRITO, 2018).

O processo de inferência da polaridade geral apresentado no módulo de AS - Nível do Documento Versão 1 considerou dois pontos: a incidência do aspecto no total geral de comentários (quanto maior a incidência maior a relevância, entendida como um peso adicionado ao aspecto) (Figura 10-2), e os aspectos de cada comentário (com sua devida polaridade) (Figura 10-3). Assim, cada aspecto positivo ou negativo de um comentário multiplica a seu valor (1 ou -1, conforme o caso) o peso do aspecto identificado anteriormente. Com a soma de todos os valores de todos os aspectos de um comentário, identifica-se a polaridade geral do comentário (Figura 10-4), que pode ser positiva ou negativa.

4.3 AS - Nível do Documento Versão 2

Nesta seção será apresentado o processo de desenvolvimento do módulo de AS - Nível do documento Versão 2. Foram realizadas as etapas apresentadas na Figura 11.

Figura 11: Etapas do processo de desenvolvimento

4.3.1. Análise das etapas

Na seção 2.4 é apresentado o trabalho de Zhang et al (2011), que descreve o processo para realizar a Análise de Sentimentos no Nível do documento na língua chinesa. Zhang et al (2011) propõem um modelo estatístico para inferir a polaridade ao documento como um todo. Uma das etapas descritas por Zhang et al (2011) é a de atribuição de pesos as sentenças que compõem o documento. Nesse sentido, Zhang et al (2011) estabelecem 5 critérios específicos para refinar a atribuição do peso das sentenças no contexto do seu documento.

A etapa de atribuição dos pesos das sentenças do modelo cascata apresentado em Zhang et al (2011) pode ser adaptada ao módulo de AS - Nível do Documento, aplicando os critérios utilizado pelo autor na definição do peso das sentenças para a definição dos pesos dos aspectos presentes nos comentários do *TripAdvisor*® analisados na *SentimentALL*. Dessa forma, o peso do aspecto no comentário não será mais definido apenas pela incidência do aspecto no total geral de comentários, como proposto por Oliveira (2018) no módulo de AS - Nível do Documento Versão 1.

4.3.2. Análise dos dados

A análise estruturada dos dados da *SentimentALL* mostrou que a forma com que os dados estavam organizados e relacionados era possível adaptar três dos critérios que Zhang et al (2011) utilizam na definição dos pesos das sentenças para a definição dos pesos dos aspectos do módulo de AS - Nível do Documento versão 1, sendo eles:

- **posição do aspecto (wp):** a posição de um aspecto em um documento pode indicar sua importância. Os aspectos situados na primeira ou última sentença do comentário são frequentemente aspectos temáticos, ou seja, o usuário tende a começar ou finalizar o

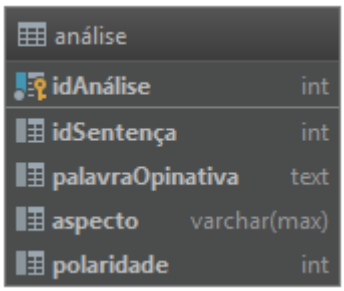
comentário falando de algo que ele mais gostou ou não, e por isso podem ser considerados mais importantes que os demais. Assim sendo, são atribuídos pesos mais altos aos aspectos que estão na primeira e última sentença do comentário.

Figura 12: Tabela sentença

A Figura 12 apresenta a estrutura da tabela sentença na base de dados da *SentimentALL*. Araújo (2017) realiza a decomposição do comentário de uma avaliação em sentenças, onde cada sentença é armazenada na base de dados com a informação da avaliação na qual ela pertence (**Figura 12 - idAvaliação**), o texto correspondente a sua respectiva parte no comentário (**Figura 12 - texto**) e sua ordem, representada pela variável **seqSentença.** Dessa forma, é possível verificar se a sentença na qual um aspecto está relacionado possui a **seqSentença** igual a 1 (início do documento/comentário) ou **seqSentença** igual a *n* (final do documento/comentário), sendo *n* a quantidade total de sentenças que o comentário possui.

- **aspectos duplicados no comentário (wd):** referem-se aos aspectos que aparecem no comentário mais de uma vez. Ou seja, a importância aumenta proporcionalmente ao número de vezes em que o aspecto aparece no comentário.

Figura 13: Tabela Análise

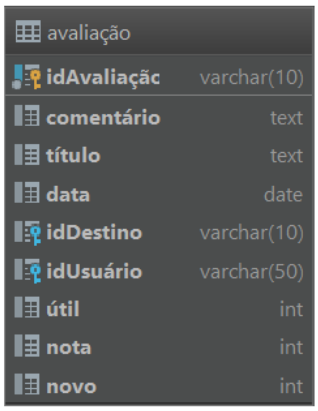


A Figura 13 apresenta a estrutura da tabela **análise** na base de dados da *SentimentALL*. Essa tabela armazena todas as análises realizadas com os aspectos em Araújo (2017) no desenvolvimento da *SentimentALL*. Dessa forma, é possível verificar se a tabela **análise** guarda mais de um registro com o mesmo aspecto (**Figura 13 - aspecto**) e sentença

(**Figura 13 - idSentença**). Esse processo deve ser realizado para cada aspecto em todas as sentenças do comentário da avaliação.

- **presença do aspecto no título (wt):** trata da presença do aspecto no título da avaliação. O título pode ser considerado um resumo do comentário, logo, um aspecto (a) presente no título (h) deve contribuir mais para o documento.

Figura 14: Tabela Avaliação



A Figura 14 apresenta a estrutura da tabela **avaliação** no banco de dados da *SentimentALL*. Dado um aspecto presente no comentário da avaliação, é possível verificar se o título (**Figura 14 - título**) contém o aspecto.

4.3.3. Alteração do banco de dados

A partir da análise da base de dados e da identificação das etapas e critérios do modelo cascata apresentado em Zhang et al (2011) possíveis de serem adaptados ao módulo de AS - Nível do Documento Versão 1, foram realizadas alterações no banco de dados para suportar a adaptação. Em vista disso, as seguintes tabelas foram criadas:

- **QtdCadaAspecto**
  - idQtdCadaAspecto
  - aspecto
  - total
- **PolaridadeGeralComentario**
  - idAvaliação
  - comentário
  - pesoFinal
  - tendênciaPolaridadeGeral
- **PesoAspecto**
  - idPesoAspecto
  - aspecto
  - pesoBase
- **AspectoAvaliacao**
  - idAspectoAvaliacao
  - idAvaliação
  - aspecto
  - polaridade

A Figura 15 apresenta o modelo relacional do banco de dados após as alterações para a adaptação do modelo cascata.

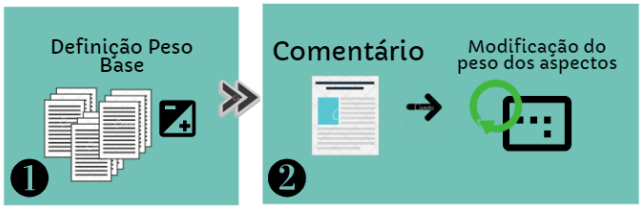
Figura 15: Modelo Relacional Módulo AS - Nível do Documento Versão 2

A Figura 15 apresenta o modelo relacional do banco de dados após as modificações para suportar a adaptação do modelo cascata. Além dos dados das tabelas **Aspectos**, **Avaliação, Análise** e **Sentença** utilizadas da *SentimentALL***,** foram criadas 4 tabelas: **QtdCadaAspecto**, **PesoAspecto**, **PolaridadeGeralComentario** e **AspectoAvaliacao**. A tabela **QtdCadaAspecto** contém os dados referentes a incidência total de cada aspecto, dado este utilizado na definição do peso base de cada aspecto. A tabela P**esoAspecto** armazena o peso base de cada aspecto. Já a estrutura **AspectoAvaliacao** relaciona de forma direta a avaliação com os aspectos analisados do seu comentário. Por fim, a tabela **PolaridadeGeralComentario** armazena o resultado da inferência da polaridade ao comentário da avaliação.

4.3.4. Módulo AS - Nível do Documento Versão 2

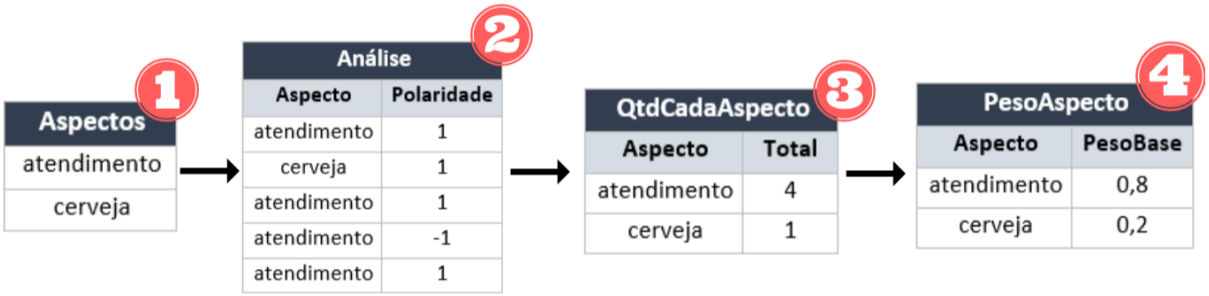
A implementação do módulo AS - Nível do Documento Versão 2 ocorreu em duas partes. A Figura 16 apresenta as etapas.

Figura 16: Etapas da implementação do Módulo AS - Nível do Documento Versão 2



A definição do **pesoBase** é dada a partir da incidência total do aspecto nos comentários (quanto maior a incidência, maior será o seu **pesoBase**). A Figura 17 apresenta de forma mais detalhada uma exemplificação da manipulação dos dados para atribuir o **pesoBase** dos aspectos.

Figura 17: Atribuição do peso base dos aspectos

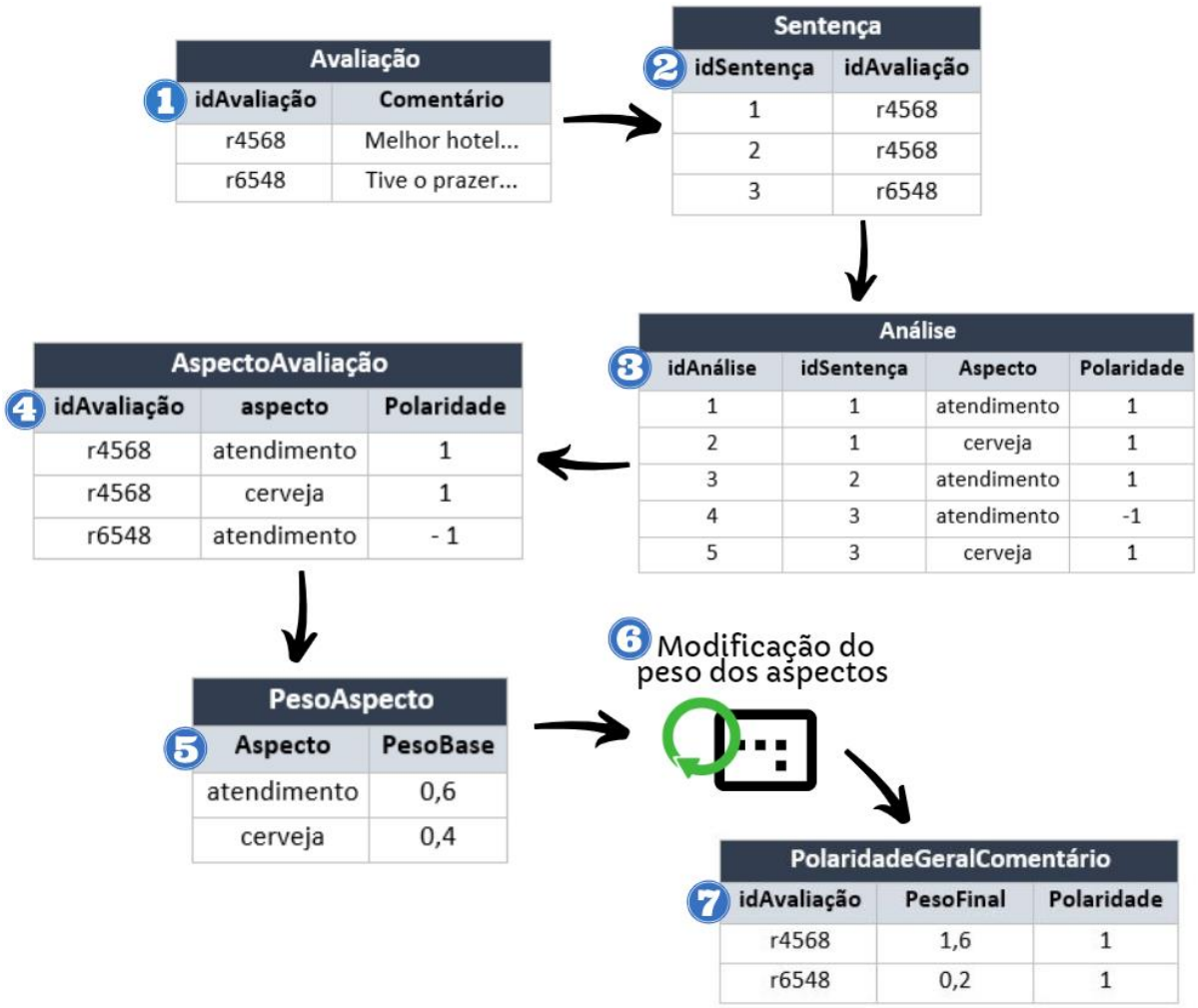


Os dados apresentados na Figura 17 tem como intuito exemplificar o processo, por esse motivo é apresentada apenas uma pequena parcela dos dados. A seguir será descrito cada passo do processo da Figura 17:

- **Aspectos** (Figura 17-1): o primeiro passo consiste em buscar os aspectos que estão contidos na tabela "**Aspectos**" na base de dados da *SentimentALL*. A tabela "**Aspectos**" possui um total de 60387 registros de aspectos diferentes. A fim de exemplificar, foi apresentado na Figura 17-1 apenas dois aspectos.
- **Análise** (Figura 17-2): em seguida, para cada aspecto, é feita uma busca na tabela "**Análise**", que possui uma coluna chamada "**Polaridade**", onde é armazenado o resultado (1 para positivo ou -1 para negativo) da análise de sentimentos do aspecto em um dado comentário efetuada pela *SentimentALL* A Figura 17-2 indica como é a estrutura dessa tabela e apresenta alguns dados sobre os aspectos "atendimento" e "cerveja". O objetivo deste processo é extrair da base de dados da *SentimentALL* (mais precisamente da tabela "**Análise**") o total geral de análise de cada aspecto, assim como o total de análises positivas e negativas de cada aspecto.
- **QtdCadaAspecto** (Figura 17-3): após verificar a incidência das análises sobre cada aspecto como descrito no passo Análise da Figura 17-2, os dados resultantes são armazenados na tabela "**QtdCadaAspecto**" para posteriormente serem usados. A Figura 17-3 ilustra o modelo da tabela com o aspecto e seu total de incidência dos aspectos atendimento e cerveja como exemplificação.
- **PesoAspecto** (Figura 17-4): a atribuição do peso base para os aspectos é feita a partir da definição da proporção entre a quantidade de análise do aspecto em particular e a quantidade de análise de todos os aspectos. Assim, o peso base é a razão da quantidade de análises do aspecto pela quantidade total de análises obtida através da soma do total de análise de cada aspecto da tabela "**QtdCadaAspecto**". Por fim, o resultado desse processamento é armazenado na tabela "**PesoAspecto**".

A segunda parte da implementação do módulo de AS - Nível do Documento Versão 2 é a inferência da polaridade ao comentário.

Figura 18: Inferência da polaridade ao comentário



A Figura 18 apresenta um exemplo do processo usado para realizar a inferência da polaridade ao comentário. Os dados apresentados representam apenas uma parcela da base de dados da *SentimentALL*. Cada etapa do processo será descrita a seguir:

- **Avaliação** (Figura 18-1): corresponde a tabela no banco de dados que possui as informações referentes às avaliações. Cada avaliação contém um comentário, que ao

final terá uma polaridade positiva ou negativa inferida. Inicialmente são recuperados os dados da avaliação.

- **Sentença** (Figura 18-2): a partir da tabela "**Sentença**" da base de dados, é possível identificar as sentenças que estão relacionadas a avaliação, sendo que cada sentença possui *n* aspectos. Cada aspecto contém uma polaridade na tabela Análise.
- **Análise** (Figura 18-3): após identificar as sentenças que estão relacionadas à avaliação, e os aspectos que pertencem às sentenças, os dados sobre o IdAvaliação, Aspecto e Polaridade de cada análise de cada aspecto presente no comentário da avaliação são armazenados na tabela "**AspectosAvaliação**".
- **AspectosAvaliação** (Figura 18-4): a tabela "**AspectosAvaliação**" relaciona os aspectos e suas polaridades com uma avaliação. Os dados aqui armazenados foram extraídos da tabela "**Análise**", que até então relacionava o aspecto a uma sentença, que por sua vez era relacionado a uma avaliação na tabela "**Sentença**".
- **PesoAspecto** (Figura 18-5): a tabela "**PesoAspecto**" contém todos os aspectos e seus respectivos pesos bases. Os pesos bases dos aspectos serão utilizados no cálculo da polaridade geral do documento.
- **Modificação dos pesos dos aspectos** (Figura 18-6): com base na abordagem cascata apresentada em Zhang et al (2011), o peso do aspecto é modificado a partir do entendimento do quão impactante e importante ele é para o comentário. Para isso, três

critérios foram estabelecidos, sendo eles: **posição do aspecto (wp)**, **aspectos duplicados no comentário (wd)** e **presença do aspecto no título (wt).**

A Figura 19 apresenta o algoritmo tendenciaPolaridadeComentario utilizado para a inferência da polaridade geral do comentário.

Figura 19: Algoritmo tendenciaPolaridadeComentario

Algoritmo 1: tendenciaPolaridadeComentario(avaliacao)

Entrada: avaliacao: uma lista com os dados de uma avaliação
Saída: tendência da polaridade do comentário

```
1: idAvaliacao ← avaliacao 0
2: comentario ← avaliacao 1
3: aspectos ← buscaAspectosAvaliação(idAvaliacao)
4: para i até |aspectos| faça:
5:     aspecto ← aspectos_{i,1}
6:     polaridade← aspectos aspectos_{i,2}
7:     pesoBase← buscaPesoBaseAspecto(aspecto)
8:     peso← pesoBase
9:     se estaNoTitulo(aspecto, idAvaliação):
10:        pesoAdicionalB ← pesoBase * 2
11:        peso← peso + pesoAdicionalB
12:    fim
13:    se aspectoRepetido(aspecto, idAvaliacao):
14:        pesoAdicionalB← pesoBase * 2
15:        peso← peso + pesoAdicionalB
16:    fim
17:    se posicaoAspecto(aspecto, idAvaliacao):
18:        pesoAdicionalC← pesoBase * 2
19:        peso← peso + pesoAdicionalC
20:    fim
22:    pesoTotal← peso * polaridade
21: fim
22: se pesoTotal > 0:
23:        tendenciaPolaridade ← 1
24: fim
25: se pesoTotal = 0:
26:        tendenciaPolaridade ← 0
27: fim
28: senão :
29:        tendenciaPolaridade ← −1
30: fim
31: retorna tendenciaPolaridade
```

O algoritmo tem como entrada uma lista contendo os dados de uma Avaliação. Nas linhas 1 e 2 o identificador da avaliação na base de dados e o comentário da avaliação são

atribuídos cada um a uma variável. Na linha 3, o algoritmo *buscaAspectosAvaliação* recebe como parâmetro o identificador da avaliação e busca no banco de dados na tabela **AspectoAvaliação** todos os registros relacionados ao identificador da avaliação. A linha 4 apresenta um *loop* para percorrer todos os aspectos que o comentário da avaliação possui. Em cada iteração do *loop* da linha 4, nas linhas 5 e 6 são atribuídos a duas variáveis distintas o aspecto e sua polaridade. A linha 7 apresenta a chamada do algoritmo *buscaPesoBaseAspecto* que recebe como parâmetro o aspecto. Esse algoritmo é responsável por buscar a informação referente ao peso base do aspecto na base de dados.

A linha 9 do algoritmo *tendenciaPolaridadeComentario* da Figura 19 apresenta o algoritmo *estaNoTitulo* que recebe como parâmetro o aspecto e o identificador da avaliação. O algoritmo é apresentado na Figura 20. Caso o retorno do algoritmo *estaNoTitulo* seja *True,* o peso base do aspecto é multiplicado por dois e somado a variável peso.

A linha 13 do algoritmo *tendenciaPolaridadeComentario* da Figura 19 apresenta a chamada do algoritmo *aspectoRepetido* que recebe como parâmetro o aspecto e o identificador da avaliação. O algoritmo é apresentado e discutido na Figura 21. Caso o retorno da função *aspectoRepetido* seja *True,* o peso base do aspecto é multiplicado por dois e somado a variável peso.

A linha 17 apresenta a chamada do algoritmo *posicaoAspecto* que recebe como parâmetro o aspecto e o identificador da avaliação. O algoritmo é apresentado na Figura 22. Caso o retorno da função *posicaoAspecto* seja *True,* o peso base do aspecto é multiplicado por dois e somado a variável peso.

Na linha 22 do algoritmo *tendenciaPolaridadeComentario* o peso do aspecto é multiplicado por sua polaridade e somado a uma variável chamada pesoTotal**,** que ao final terá o peso final do comentário. Se o pesoTotal for maior que zero, é inferida a polaridade positiva (1) ao comentário. Se for igual a zero, é inferida a polaridade neutra (0) ao comentário. Por fim, se o pesoTotal for menor que zero, é inferida a polaridade negativa (-1) ao comentário da avaliação. Ao final, o resultado com a polaridade do comentário é retornado.

Figura 20: Algoritmo estaNoTitulo

Algoritmo 2: estaNoTitulo (aspecto, idAvaliacao)

Entrada: aspecto: aspecto a ser verificado se esta no título, idAvaliacao: identificador da avaliação na base de dado
Saída: valor booleano

```
1: avaliacao← buscaAvaliacao(idAvaliacao)
2: titulo← avaliação 2
3: se aspecto está no titulo:
4:       retorne True
5: senão:
6:       retorne False
7: fim
```

O algoritmo *estaNoTitulo* apresentado na Figura 20 recebe como entrada um aspecto e o identificador da avaliação. Na linha 1 do algoritmo *estaNoTitulo*, o algoritmo *buscaAavaliacao* busca a avaliação no banco de dados a partir do identificador recebido como parâmetro e retorna os dados da avaliação no formato de lista. A linha 3 do algoritmo *estaNoTitulo* é responsável por verificar se o aspecto está no título da avaliação. Caso esteja, é retornado o valor *booleano True.* Caso não esteja, é retornando o valor *booleano False.*

Figura 21: Algoritmo aspectoRepetido

Algoritmo 3: aspectoRepetido(aspecto, idAvaliacao)

Entrada: aspecto: aspecto a ser verificado se esta repetido no comentário, idAvaliacao: identificador da avaliação na base de dado
Saída: valor booleano

```
1: analisesAspecto ← buscaAnalisesAspectoDaAvaliacao(aspecto, idAvaliacao)
2: se |analisesAspecto| > 1:
3: |     retorne True
5: senão:
6: |     retorne False
7: fim
```

O algoritmo *aspectoRepetido* apresentado na Figura 21 recebe como entrada um aspecto e o identificador da avaliação. Na linha 1 do algoritmo *aspectoRepetido*, o algoritmo *buscaAnalisesAspectoDaAvaliacao* busca todas as análises do aspecto em uma avaliação na base de dados. A linha 2 do algoritmo 3 é responsável por verificar se a quantidade de análises de um aspecto é maior do que um. Caso seja, é entendido que o aspecto está presente mais de uma vez no comentário da avaliação, e então, o valor *booleano True* é retornado. Caso a quantidade de análises seja um, é retornando o valor *booleano False.*

Figura 22: Algoritmo posicaoAspecto

Algoritmo 4: posicaoAspecto(aspecto, idAvaliacao)

Entrada: aspecto: aspecto para verificar sua posição no documento, idAvaliacao: identificador da avaliação na base de dado
Saída: valor booleano

1: sentencasAvaliacao← $buscaSentencasDaAvaliacao(idAvaliacao)$
2: n ← $|sentencasAvaliacao| - 1$
3: **para** i **até** n **faça:**
4: sentenca ← $sentencasAvaliacao_i$
5: **se** $aspecto \in ((sentenca_3) \wedge (sentenca_2 = 0 \vee sentenca_2 = n))$:
6: **retorne** $True$
7: **fim**
8: **retorne** $False$
9: **fim**

O algoritmo apresentado na Figura 22 recebe como entrada um aspecto e o identificador da avaliação. Na linha 1 do algoritmo *posicaoAspecto*, o algoritmo *buscaSentencasDaAavaliacao* busca todas as sentenças do comentário da avaliação no banco de dados a partir do identificador recebido como parâmetro. O objetivo do algoritmo é verificar em qual sentença do comentário o aspecto está presente. Caso o aspecto esteja na primeira ou na última sentença do comentário, o valor *booleano True* é retornado. Caso o aspecto não esteja na primeira ou última sentença do comentário, o valor *booleano False* é retornado.

Ao final do processo de modificação dos pesos dos aspectos e da inferência da polaridade ao comentário, esses dados são armazenados no banco de dados na tabela **polaridadeGeralComentario.**

4.4 Avaliação da qualidade do resultado

Para avaliar a qualidade do resultado do módulo de AS - Nível do Documento Versão 1 e o módulo de AS - Nível do Documento versão 2 desenvolvido neste trabalho foram aplicadas as métricas *precision, recall* e *F-measure*.

A seguir é apresentado um exemplo prático da aplicação do processo a fim de auxiliar no entendimento das etapas para a realização da análise.

A primeira etapa do processo consiste na seleção dos comentários a serem analisados. Tendo como entrada a base de dados da *SentimentALL*, um algoritmo é executado e de forma aleatória para selecionar uma avaliação na tabela **avaliação** no banco de dados. A informação referente ao identificador da avaliação no banco de dados, o texto do comentário e seu valor na

escala *Likert* são armazenados em um arquivo CSV para mais tarde serem utilizados na análise manual.

Tabela 7: Avaliações selecionadas

Id	Texto do Comentário	Escala Likert
r100000545	Fugindo do calor fomos para a serra e nos hospedamos pela no Rosa dos Ventos, foi tudo muito bom, das acomodações com uma gastronomia sempre perfeita. Para lamentar, somente a chuva, muita chuva durante 5 dias, o que nos impediu que aproveitássemos ainda mais. Não passeamos pelas trilhas, não vimos a vista nem o mirante, mas fomos bem cuidados pela equipe e compensados pelo congraçamento espontâneo dos hóspedes no Bar Anglais e jantares temáticos, parecia uma família. Apesar da chuva tivemos um Carnaval Nota 10. Vamos voltar	5
r100043225	Esse restaurante nunca decepciona em termos de comida, serviço e preço!!! O serviço é atencioso e a comida deliciosa. Recomendação especial para os pratos de inverno, tipo Fondue e Raclete, que são deliciosos !!!!	4

A Tabela 7 apresenta uma amostra dos dados após a seleção e o armazenamento no arquivo CSV.

A partir do arquivo CSV contendo as avaliações selecionadas, é realizado o processo de análise manual, onde no mesmo arquivo é acrescentada uma coluna com a informação do resultado da análise manual de cada comentário.

Tabela 8: Análise manual

Id	Texto do Comentário	Escala Likert	Análise Manual
r100000545	Fugindo do calor fomos para a serra e nos hospedamos pela no Rosa dos Ventos, foi tudo muito bom, das acomodações com uma gastronomia sempre perfeita. Para lamentar , somente a chuva, muita chuva durante 5 dias, o que nos impediu que aproveitássemos ainda	5	1

	mais. Não passeamos pelas trilhas, não vimos a vista nem o mirante, mas fomos bem cuidados pela equipe e compensados pelo congraçamento espontâneo dos hóspedes no Bar Anglais e jantares temáticos, parecia uma família. Apesar da chuva tivemos um Carnaval Nota 10. Vamos voltar !!!		
r100043225	Esse restaurante nunca decepciona em termos de comida, serviço e preço!!! O serviço é atencioso e a comida deliciosa. Recomendação especial para os pratos de inverno, tipo Fondue e Raclete, que são deliciosos !!!!	4	1

A Tabela 8 apresenta uma amostra dos dados após a análise manual. No processo de análise manual, cada comentário foi classificado com o valor 1 (para o comentário entendido como positivo), -1 (para o comentário entendido como negativo) ou 0 (neutro), caso não fosse possível entender o comentário como positivo ou negativo.

Após a análise manual, a próxima etapa foi de inferência da polaridade geral do sistema aos comentários selecionados e analisados manualmente. Para isso foram realizados dois passos:

- recuperação das avaliações selecionadas utilizando os Id's armazenados no arquivo CSV;
- inferência da polaridade geral do comentário pelo sistema das avaliações selecionadas.

Após a polaridade dos comentários das avaliações também serem inferidas pelo sistema, as informações acerca do identificador da avaliação, resultado da análise manual e da análise do sistema, além do valor do comentário na escala *Likert,* foram organizadas em uma única planilha.

Tabela 9: Análise do Sistema

Id	Análise Manual	Análise do Sistema	Escala Likert
r100000545	1	1	5
r100043225	1	1	4

A Tabela 9 apresenta uma amostra dos resultados da análise manual, análise do sistema, identificador da avaliação e seu valor na escala *Likert* organizados em uma única tabela para mais tarde serem utilizados na criação da matriz de confusão e análise de correlação.

As avaliações compreendidas como neutras (valor 0) na análise manual ou do sistema foram desconsideradas. Em seguida, foram aplicadas as métricas *precision, recall* e *F-measure.*

A Tabela 10 apresenta o resultado obtido ao aplicar o processo de avaliação da qualidade do resultado do módulo de AS - Nível do Documento Versão 1.

Tabela 10: Aplicação da Matriz de Confusão módulo AS - Nível do Documento Versão 1

#	Positivo	Negativo
Verdadeiro	87	
Falso	10	3

Os campos presentes na matriz de confusão da Tabela 10 são referentes a análise manual e a análise realizada pelo módulo de AS - Nível do Documento versão 1. A análise manual e a análise no sistema identificaram de forma equivalente (Verdadeiro - Positivo) a polaridade de um total de 87 comentários. Já o total de comentários em que o sistema classificou com a polaridade negativa e a análise manual classificou como positiva foram 3. Os comentários entendidos com a polaridade negativa na análise manual e com polaridade positiva na análise do sistema totalizaram 10 comentários. Portanto, ao aplicar as métricas de *precision, recall* e *F-measure*, tem-se que $precision = \frac{87}{87 + 10}$, cujo resultado é 0,896. *Precision* fornece informação sobre Falsos - Positivos, então trata-se de quão bom o modelo é para identificar o resultado de maneira precisa, isto é, dos classificados como positivos, quantos realmente são. Logo, quando mais próximo de 1 for o resultado, significa que menor é a possibilidade de o modelo classificar como positivo um comentário que não é considerado positivo na análise manual.

A representação da métrica *Recall* é obtida a partir de $recall = \frac{87}{87 + 3}$, cujo resultado é 0,966. *Recall* indica a relação entre as polaridades positivas corretamente identificadas pelo sistema (considerando a avaliação manual) e todas as previsões que realmente são positivas (ou seja, o conjunto *True Positive* + *False Negative* ou a proporção dos positivos identificados corretamente). Portanto quanto mais próximo o resultado de *Recall* for de 1, melhor será o modelo para identificar os comentários positivos.

Para mensurar o avaliação geral do módulo foi utilizada a métrica *F-Measure*, definida por $F - Measure = 2 * \frac{0,896 * 0,966}{0,896 + 0,966}$, gerando o resultado de 0,929. Essa é a média harmônica entre *precision* e *recall*, e seu valor máximo é 1. Com *F-Measure*, quanto mais próximo do valor 1, melhor a avaliação do módulo.

A Tabela 11 apresenta o resultado obtido ao aplicar o processo de avaliação da qualidade do resultado do módulo de AS - Nível do Documento versão 2.

Tabela 11: Aplicação da Matriz de Confusão módulo AS - Nível do Documento Versão 2

#	Positivo	Negativo
Verdadeiro	89	
Falso	9	2

Ao aplicar as métricas de *precision*, *recall* e *F-measure* da mesma forma que foi aplicada ao resultado do módulo de AS - Nível do Documento versão 1, tem-se que$precision = \frac{89}{89 + 9}$, cujo resultado é 0,908, $recall = \frac{89}{89 + 2}$, sendo o resultado 0,978 e $F - Measure = 2 * \frac{0,908 * 0,978}{0,908 + 0,978}$, gerando o 0,941.

4.5 Processo de correlação

Para verificar qual a relação existente entre o resultado obtido nos módulos de AS-Nível do Documento Versão 1 e Versão 2 com os dados referentes a avaliação via escala *Likert* feita pelos usuários do *TripAdvisor*®, foi realizada a análise de correlação de *Pearson* utilizando o Excel.

Tabela 12: Escala Likert

#	ID AVALIAÇÃO	ANÁLISE DO SISTEMA	ESCALA LIKERT
1	r100000545	1	5
2	r100016264	1	4
3	r100080466	-1	2
4	r100043225	1	5

A Tabela 12 apresenta uma amostra dos dados utilizados para a análise de correlação. Id Avaliação é o identificador da avaliação na base de dados da *SentimentALL*, análise do sistema é o resultado alcançado com o módulo de AS - Nível do Documento, o valor na escala *Likert* é o expressado pelo usuário recuperado da base de dados da *SentimentALL*. Para a análise de correlação, os valores expressos na escala *Likert* foram compreendidos de uma forma diferente. Para isso, foi considerado o seguinte critério: 1 e 2 (na escala *Likert*) = -1 (para a análise de correlação), 3 (na escala *Likert*) foi desconsiderado para a análise de correlação, 4 e 5 (na escala *Likert*) = 1 (para a análise de correlação). Desta forma, o valor da escala *Likert* de cada avaliação foi atualizado com o valor utilizado pelo módulo de identificação da polaridade geral do comentário da *SentimentALL*, que trabalha com 1 (para positivo) e -1 (para negativo).

Tabela 13: Dados para análise de correlação

#	ID AVALIAÇÃO	ANÁLISE DO SISTEMA	ESCALA LIKERT
1	r100000545	1	1
2	r100016264	1	1
3	r100080466	-1	-1
4	r100043225	1	1

A Tabela 13 apresenta os dados preparados para a análise de correlação. A coluna polaridade contém o resultado da inferência da polaridade feita pelo sistema. Os valores da escala *Likert* estão expressos de forma equivalente ao resultado obtido pelo sistema, conforme explicação anterior.

Nesse sentido, a correlação colabora no entendimento de que o módulo está ou não apresentando resultados de fato significativos na inferência correta da polaridade a um comentário. Isso porque verifica se o valor na escala *Likert* dada pelo usuário em uma determinada avaliação possui ou não uma correlação direta com o seu texto em forma de comentário na mesma avaliação. E o coeficiente de correlação mostra se essa correlação existe de forma consistente ou não.

Tabela 14: Dados para análise de correlação do módulo AS - Nível do Documento versão-1

#.	ID AVALI	AS - V1	LIKERT	#.	ID AVALI	AS - V1	LIKERT
1	r100000545	1	1	27	r100212285	1	1
2	r100016264	-1	1	28	r100216188	1	1
3	r100030422	1	1	29	r100216692	1	1
4	r100043225	1	1	30	r100217410	1	1
5	r100051235	1	1	31	r100261730	1	1
6	r100051709	1	-1	32	r100262747	-1	-1
7	r100078074	1	1	33	r100266945	1	1
8	r100080466	-1	-1	34	r100272442	-1	-1
9	r100083167	-1	-1	35	r100272466	1	1
10	r100108630	1	1	36	r100277235	1	1
11	r100120511	1	1	37	r100293130	1	1
12	r100122191	1	1	38	r100325495	-1	-1
13	r100135625	1	1	39	r100365140	1	1
14	r100157458	1	1	40	r100371377	1	1
15	r100160671	1	1	41	r100371856	-1	1
16	r100161573	1	1	42	r100397989	1	1
17	r100165846	1	-1	43	r100410747	-1	-1
18	r100170847	1	1	44	r100431058	1	1
19	r100178645	1	1	45	r100432458	1	1
20	r100182253	1	1	46	r100434833	1	1
21	r100197410	1	1	47	r100445288	1	1
22	r100200658	1	1	48	r100480278	1	1
23	r100207271	1	1	49	r100488011	-1	-1
24	r100208063	1	1	50	r100493981	1	1
25	r100209906	1	1	51	r100494876	-1	-1
26	r100210361	1	1	52	r100514302	1	1
53	r100212285	1	1	78	r100752611	1	1

54	r100216188	1	1	79	r100754322	1	1
55	r100216692	1	1	80	r100754909	1	-1
56	r100217410	1	1	81	r100758099	1	1
57	r100261730	1	1	82	r100764629	1	1
58	r100262747	-1	-1	83	r100766312	1	1
59	r100266945	1	1	84	r100777314	1	1
60	r100272442	-1	-1	85	r100794564	-1	-1
61	r100272466	1	1	86	r100809957	1	1
62	r100559946	1	1	87	r100828525	1	-1
63	r100560019	1	1	88	r100844774	-1	-1
64	r100560159	1	1	89	r100847008	1	1
65	r100561147	-1	-1	90	r100847502	1	1
66	r100564256	1	1	91	r100851700	1	1
67	r100606041	1	1	92	r100871388	1	1
68	r100613174	1	1	93	r100873328	1	1
69	r100622369	1	1	94	r100875347	1	1
70	r100649054	1	1	95	r100878597	1	-1
71	r100671812	1	1	96	r100879135	1	1
72	r100672025	1	1	97	r100889727	1	1
74	r100707266	1	-1	98	r100895461	1	1
75	r100720046	1	1	99	r100927524	1	-1
76	r100731376	1	1	100	r100560159	1	1
77	r100737886	1	1				

A Tabela 14 apresenta os dados preparados para a análise de correlação do módulo AS - Nível do Documento Versão 1. A coluna AS-1 contém o resultado da inferência da polaridade feita pelo sistema. O resultado obtido após a análise de correlação de *Pearson* foi o coeficiente de correlação igual a 0,6621. Esse resultado será explicado posteriormente.

Tabela 15: Dados para análise de correlação do módulo AS - Nível do Documento versão-2

#.	ID AVALI	AS - V2	LIKERT	#.	ID AVALI	AS - V2	LIKERT
1	r100000545	1	1	27	r100212285	1	1
2	r100016264	-1	1	28	r100216188	1	1
3	r100030422	1	1	29	r100216692	1	1
4	r100043225	1	1	30	r100217410	1	1
5	r100051235	1	1	31	r100261730	1	1
6	r100051709	1	-1	32	r100262747	-1	-1
7	r100078074	1	1	33	r100266945	1	1
8	r100080466	-1	-1	34	r100272442	-1	-1
9	r100083167	-1	-1	35	r100272466	1	1
10	r100108630	1	1	36	r100277235	1	1
11	r100120511	1	1	37	r100293130	1	1
12	r100122191	1	1	38	r100325495	-1	-1
13	r100135625	1	1	39	r100365140	1	1
14	r100157458	1	1	40	r100371377	1	1
15	r100160671	1	1	41	r100371856	-1	1
16	r100161573	1	1	42	r100397989	1	1
17	r100165846	1	-1	43	r100410747	-1	-1
18	r100170847	1	1	44	r100431058	1	1
19	r100178645	1	1	45	r100432458	1	1
20	r100182253	1	1	46	r100434833	1	1
21	r100197410	1	1	47	r100445288	1	1
22	r100200658	1	1	48	r100480278	1	1
23	r100207271	1	1	49	r100488011	-1	-1
24	r100208063	1	1	50	r100493981	1	1
25	r100209906	1	1	51	r100494876	-1	-1
26	r100210361	1	1	52	r100514302	1	1
53	r100524713	1	1	77	r100737886	1	1
54	r100525340	1	1	78	r100752611	1	1

55	r100527027	1	1	79	r100754322	1	1
56	r100527282	-1	-1	80	r100754909	1	-1
57	r100528658	1	1	81	r100758099	1	1
58	r100538700	1	1	82	r100764629	1	1
59	r100549527	1	-1	83	r100766312	1	1
60	r100556156	1	1	84	r100777314	1	1
61	r100557857	1	1	85	r100794564	-1	-1
62	r100559946	1	1	86	r100809957	1	1
63	r100560019	1	1	87	r100828525	1	-1
64	r100560159	1	1	88	r100844774	-1	-1
65	r100561147	-1	-1	89	r100847008	1	1
66	r100564256	1	1	90	r100847502	1	1
67	r100606041	1	1	91	r100851700	1	1
68	r100613174	1	1	92	r100871388	1	1
69	r100622369	1	1	93	r100873328	1	1
70	r100649054	1	1	94	r100875347	1	1
71	r100671812	1	1	95	r100878597	1	-1
72	r100672025	1	1	96	r100879135	1	1
73	r100672266	1	1	97	r100889727	1	1
74	r100967263	1	1	98	r100895461	1	1
75	r100720046	1	1	99	r100927524	1	-1
76	r100731376	1	1	100	r100946191	1	1

A Tabela 15 apresenta os dados preparados para a análise de correlação do módulo AS - Nível do Documento Versão 2. Seguindo o mesmo modelo da organização dos dados na análise de correlação do módulo versão 1, a coluna AS-2 contém o resultado da inferência da polaridade feita pelo sistema. O resultado obtido após a análise de correlação de *Pearson* foi o coeficiente de correlação igual a 0,6861.

Lira (2004, p. 41) coloca que o coeficiente de correlação entre "$0{,}60 \leq \rho < 0{,}90$, existe forte correlação linear;". Por esse motivo, é possível dizer que ambos os módulos alcançaram bons resultados. Isso porque considerou que o valor na escala *Likert* dada pelo

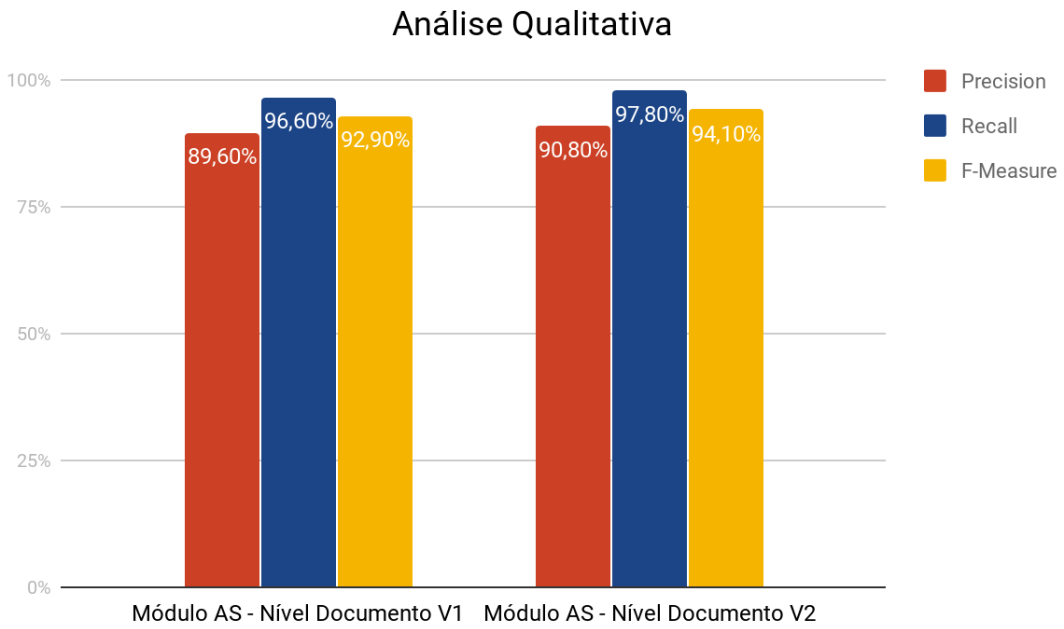
usuário em uma determinada avaliação possui uma correlação direta com o seu texto em forma de comentário na mesma avaliação. É necessário ressaltar que a análise de correlação foi realizada considerando apenas 100 avaliações, o que representa uma pequena parcela das mais de quatro milhões de avaliações presentes na base de dados da *SentimentALL*. Logo, para apresentar um resultado de forma mais concreta, é preciso que a análise de correlação seja feita considerando uma amostra de dados mais abrangente.

Na seção a seguir será apresentado um paralelo entre os resultados alcançados nas versões 1 e 2 do módulo de AS - Nível do Documento.

4.6 Paralelo do resultado - versão 1 e versão 2

Nesta seção será apresentada uma comparação entre os resultado alcançados pelo módulo de AS - Nível do Documento versão 1 e versão 2. O objetivo é mostrar qual obteve o melhor resultado na classificação da polaridade dos comentários da base de dados da *SentimentALL*. A Figura 23 apresenta um gráfico com os resultados obtidos na aplicação das métricas *precision*, *recall* e *F-measure* no módulo versão 1 e versão 2.

Figura 23: Análise dos resultados

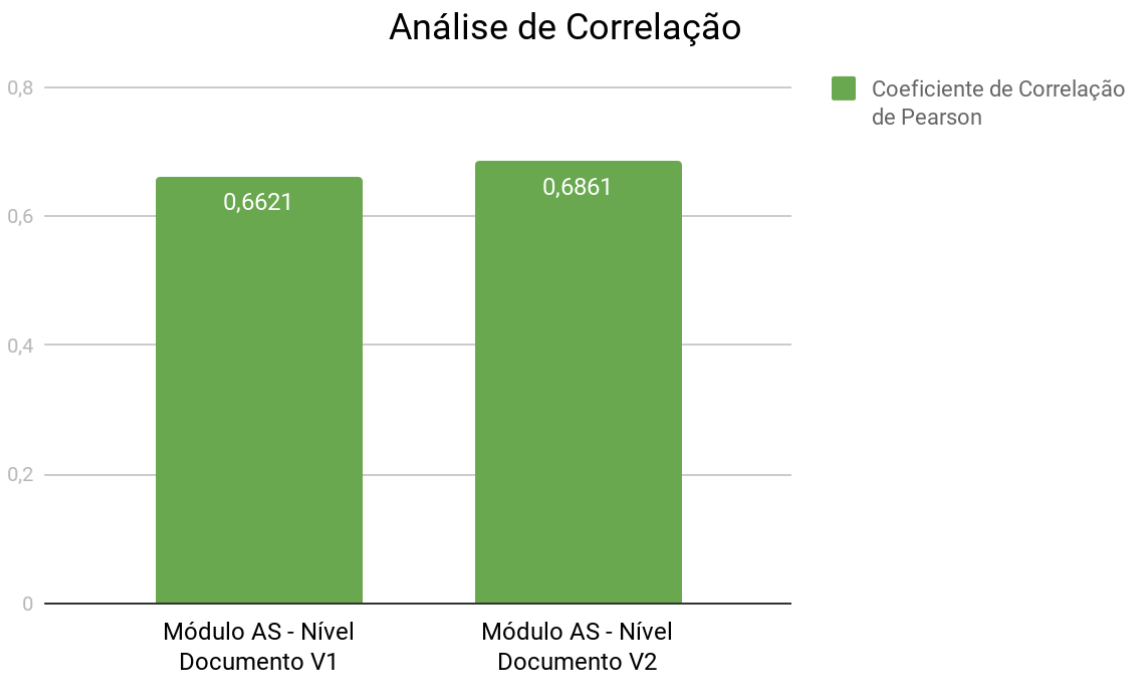


Como pode ser observado na Figura 23, o módulo de AS - Nível do Documento versão 2 desenvolvido neste trabalho alcançou um resultado melhor nas três métricas aplicadas para verificar a qualidade do resultado em comparação ao módulo de AS - Nível do Documento versão 1. *Precision* obteve 90,80%, tendo uma diferença de 1,2% em relação ao o módulo de

AS - Nível do Documento versão 1, que obteve um resultado de 89,60%. Em *recall*, o resultado alcançado no módulo versão 2 foi de 97,80%, já o módulo da versão 1 obteve o resultado equivalente a 96,60%.

Por fim, *F-measure* que representa uma medida harmônica entre as duas métricas *precision* e *recall* chegou a marca de 94,10% para o módulo de AS - Nível do Documento versão 2, e 92,90% para o módulo de AS - Nível do Documento versão 1.

Figura 24: Análise de correlação



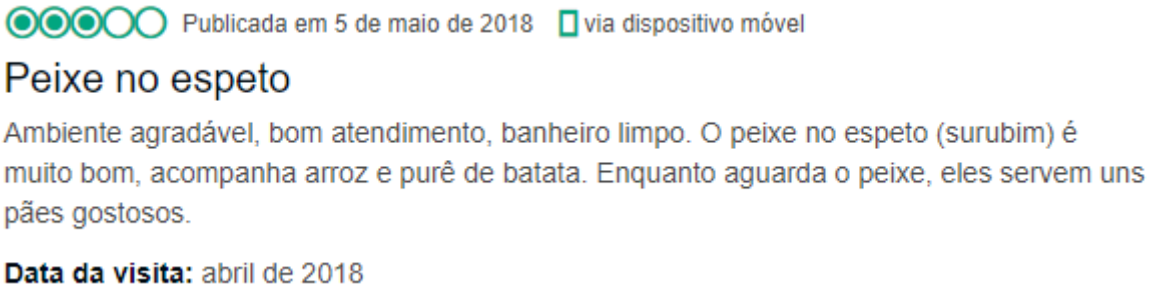
A análise de correlação corroborou com o entendimento de que o módulo está apresentando bons resultados. Isso porque considerou que o valor na escala *Likert* dada pelo usuário em uma determinada avaliação possui uma correlação direta com o seu texto em forma de comentário na mesma avaliação. Lira (2004, p. 41) aponta que o coeficiente de correlação entre "$0{,}60 \leq \rho < 0{,}90$, existe forte correlação linear;". O módulo de AS - Nível do Documento versão 2 apresentou o coeficiente de correlação igual a 0,6861. Já o módulo de AS - Nível do Documento versão 1 apresentou o coeficiente de correlação igual a 0,6621, como apresentado na Figura 24.

Dessa forma, a conclusão obtida é a de que o módulo de AS - Nível do Documento versão 2 apresentou melhores resultados em relação ao módulo de AS - Nível do Documento

versão 1, inferindo assertivamente a polaridade geral (positiva ou negativa) a um número maior de comentários, considerando os mesmos dados.

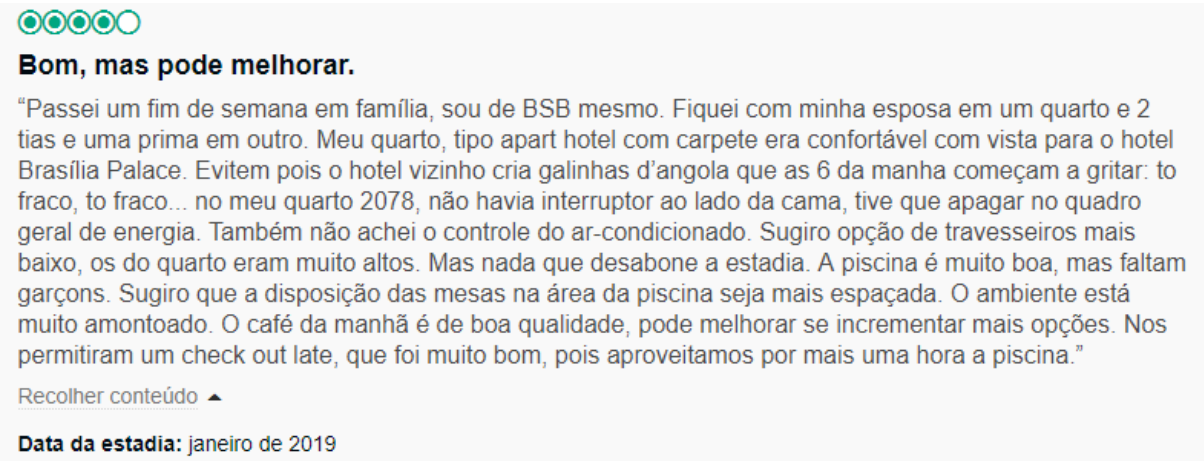
É necessário ressaltar que alguns fatores podem influenciar no resultado da análise de correlação, como por exemplo, o usuário expressar uma opinião positiva no comentário e avaliar com um valor baixo (1, 2 ou 3) via escala *Likert*, ou expressar uma opinião negativa no comentário e avaliar com uma nota alta (4 ou 5) via escala *Likert.* A Figura 25 apresenta um exemplo de uma avaliação do *TripAdvisor* onde o usuário da nota 3 na escala *Likert*, porém, relata uma ótima experiência no comentário.

Figura 25: Avaliação do TripAdvisor - 01



Do mesmo modo, a Figura 26 apresenta uma avaliação onde o usuário avaliou com valor 4 na escala *Likert*, porém o texto do seu comentário pode ser classificado como negativo.

Figura 26: Avaliação do TripAdvisor - 02



Apesar da diferença da qualidade do resultado apresentado entre os módulos versão 1 e 2 na aplicação das métricas ter sido pequena, é importante ressaltar que os testes de qualidade foram realizados considerando 100 avaliações, o que representa uma pequena parcela do total de mais de 4 milhões de avaliações analisadas pela *SentimentALL*. Por esse motivo, o pequeno

percentual de diferença apresentado nos testes considerando 100 avaliações, poderá representar uma diferença mais significativa na quantidade de comentários classificados corretamente na AS - Nível do Documento aplicada a todas as avaliações da base de dados da *SentimentALL*.

Essa seção apresentou os resultados obtidos no desenvolvimento do módulo de AS - Nível do Documento versão 2, da avaliação da qualidade do resultado do módulo de AS - Nível do Documento versão 1 e 2 e uma comparação dos resultados da aplicação das métricas *precision, recall* e *F-measure* nos módulos versão 1 e versão 2.

A seção a seguir será apresentado as considerações finais bem como os possíveis trabalhos futuros

5. CONSIDERAÇÕES FINAIS

Este estudo foi concentrado em aplicar uma adaptação da abordagem cascata apresentada em Zhang et al (2011) ao módulo de AS - Nível do Documento versão 1, desenvolvendo o módulo de AS - Nível do Documento versão 2. Além disso, foi analisada e comparada a qualidade dos resultados obtidos entre os módulos versão 1 e versão. Para isso, foram realizadas quatro etapas, sendo elas a análise das etapas, análise dos dados, alterações no banco e implementação.

A análise das etapas consistiu em compreender o modelo cascata apresentado por Zhang et al (2011). A partir desse entendimento, foi realizada uma análise dos dados da *SentimentALL* a fim de verificar quais etapas do modelo cascata eram possíveis de serem adaptadas. Em seguida, foram feitas alterações no banco de dados para suportar a adaptação do modelo cascata.

O contexto em que Zhang et al (2011) apresentaram o modelo cascata foi para a definição da polaridade de documentos e artigos escrito na língua chinesa. Os autores explicaram que a língua chinesa apresenta maiores desafios para a AS devido a sua complexidade na identificação de alguns termos. Eles apresentaram outras técnicas utilizadas junto a abordagem cascata para a identificação das sentenças do documento, definição das palavras chaves, além de outros critérios para definição do peso da sentença de acordo com o contexto do documento, que não foram adaptados ao módulo de AS - Nível do Documento. Por exemplo, realizar a definição de pesos das sentenças com base nos aspectos presentes na mesma e, a partir disso, utilizar os pesos das sentenças para realizar a inferência da polaridade ao documento.

A partir disso, a reestruturação do módulo de AS - Nível do Documento versão 1 para o módulo de AS - Nível do Documento versão 2 consistiu na modificação da atribuição do peso aos aspectos presentes no comentário. Cada aspecto recebeu um peso base e outros três critérios (verificação se o aspecto está presente no título da avaliação, repetição do aspecto no comentário e a posição do aspecto no comentário) foram estabelecidos para refinar o entendimento do quão importante é o aspecto para o contexto do documento no momento da inferência da sua polaridade geral. Por meio da aplicação das métricas de *precision*, *recall* e *F-measure* foi verificado que houve melhorias na capacidade do módulo de inferir corretamente a polaridade ao comentário em relação ao módulo de AS - Nível do Documento versão 1.

Na avaliação da qualidade dos resultados dos módulos de AS - Nível do Documento versão 1 e versão 2 foram aplicadas as métricas *precision*, *recall* e *F-measure* junto com a estratégia de organização dos dados chamada matriz de confusão.

Ao comparar os resultados, o módulo de AS - Nível do Documento versão apresenta melhorias em relação ao módulo de AS - Nível do Documento, mostrando que a adaptação da abordagem cascata utilizada para refinar o entendimento do quão impactante e importante o aspecto é para o comentário foi válida.

Apesar da análise de correlação ter mostrado que o coeficiente de correlação obtido como resultado é considerado como forte, Lira (2004) diz que alguns fatores podem influenciar na intensidade do coeficiente de correlação, como, por exemplo, o tamanho da amostra utilizada para análise. Além disso, é necessário ressaltar que existe a possibilidade de o usuário expressar uma opinião no comentário de forma diferente do valor expresso na escala *Likert.*

Considerando esses fatores, a partir dessa pesquisa, podem ser desenvolvidos trabalhos futuros com o intuito de aplicar análise de correlação de *pearson* e as métricas *precision*, *recall* e *F-measure*, a partir de uma amostra de dados maior. No entanto, o cenário ideal é realizar a AS - Nível do Documento e o processo de análise da qualidade do resultado considerando todas as avaliações do banco da *SentimentALL*. Todavia, para isso é necessário que haja um tempo e processamento computacional maior, em razão da base de dados da *SentimentALL* possuir mais de 4 milhões de avaliações.

Ainda para trabalhos futuros, outras técnicas para a AS - Nível de Documento podem ser desenvolvidas, por exemplo, aprendizagem de máquina, que incluem *Naive Bayes*, *Support Vector Machines* (SVM) como apresenta Bibi (2017).

Além disso, também há necessidade do estudo de outras métricas para a avaliação da qualidade do resultado, como *Kappa* apresentado em Figueiredo e Vieira (2007) e ROC (Braga, 2001). Outra possibilidade de trabalho futuro é realizar a integração do módulo de AS - Nível do Documento versão 2 à ferramenta *SentimentALL*.

REFERÊNCIAS

ARAÚJO, Luan Gomes de Almeida. **SENTIMENTALL VERSÃO 2: Desenvolvimento de Análise de Sentimentos em Python.** 2017. 125f. TCC II (Bacharel em Ciência da Computação) - Ceulp Ulbra, Palmas.

ALBORNOZ, Jorge Carrillo de; PLAZA, Laura; GERVAS, Pablo; DIAZ, Alberto. **A joint model of feature mining and sentiment analysis for product review rating**. In: European conference on information retrieval. Springer, Berlin, Heidelberg, 2011. p. 55-66. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-642-20161-5_8>. Acesso em 16 de set de 2018.

BOLLEN, J., Mao, H., and Zeng, (2010). **Twitter mood predicts the stock market**. CoRR, abs/1010.3003. Disponível em: <https://arxiv.org/pdf/1010.3003.pdf>. Acesso em: 20 de set de 2018.

BENEVENUTO, Fabrício; RIBEIRO, Filipe; ARAÚJO, Matheus. **Métodos para Análise de Sentimentos em mídias sociais**. 2015. 30 f. Minicurso (Minicurso) - x, [S.l.], 2015. Disponível em: <http://homepages.dcc.ufmg.br/~fabricio/download/webmedia-short-course.pdf>. Acesso em 16 de set de 2018.

BAKER, A. **Computer Aided Invariant Feature Selection**. (Doctoral Dissertation). University of Florida. Retrieved from. Disponível em: <http://ufdc.ufl.edu/UFE0022870>. Acesso em: 20 de abril de 2019.

BRITO, Parcilene Fernandes de. **RELATOS VERBAIS DE CONSUMIDORES EM AVALIAÇÕES ON-LINE:** PROSPECÇÃO COMPUTACIONAL E INTERPRETAÇÕES COM BASE NO BEHAVIORAL PERSPECTIVE MODEL (BPM). 2018. 182 f. Tese (Programa de Pós-Graduação STRICTO SENSU em Psicologia) - Pontifícia Universidade Católica de Goiás, Goiânia-GO.

BECKER, Karin; TUMITAN, Diego**. Introdução à Mineração de Opiniões: Conceitos, Aplicações e Desafio**s. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 28, 2013, Recife. Minicursos. Recife: Sbbd, 2013. Disponível em: <http://inf.ufrgs.br/~kbecker/lib/exe/fetch.php?media=minicursosbbd_versaosubmetida.pdf>. Acesso em: 28 de nov. 2018.

BLEI, D.M., Ng, A.Y. and Jordan, M.I. (2003) **Latent Dirichlet Allocation**. The Journal of Machine Learning Research, 3, 993-1022. Disponível em: <http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>. Acesso em: 28 nov 2018.

BRAGA, Ana Cristina da Silva. **CURVAS ROC: ASPECTOS FUNCIONAIS E APLICAÇÕES.** Dissertação submetida à Universidade do Minho para obtenção do grau de doutor - Universidade do Minho. 267 f. Braga - Portugal. Disponível em: <http://repositorium.sdum.uminho.pt/bitstream/1822/195/1/tese_doutACB.pdf>. Acesso em: 9 de jun. 2019.

BRITO, Edeleon Marcelo. Mineração de Textos: **Detecção automática de sentimentos em comentários nas mídias sociais**. 2016. Projeto de Dissertação de Mestrado Profissional em

Sistemas da Informação e Gestão do Conhecimento) — Universidade Fundação Mineira de Educação Cultura. Belo Horizonte, 2016. Disponível em: <http://www.fumec.br/revistas/sigc/article/download/3737/2034>. Acesso em 23 de set de 2018.

BRITO, P. F.. SentimentALL: **Ferramenta de Análise de Sentimentos em Língua Portuguesa com foco no Setor do Turismo**. Grupo de Pesquisa Engenharia Inteligente de Dados: CEULP/ULBRA, 2015.

CAMBRIA, Erik; DAS, Dipankar; BANDYOPADHYAY, Sivaji; FERACO, Antonio. Affective Computing and Sentiment Analysis. In: CAMBRIA, Erik; DAS, Dipankar; BANDYOPADHYAY, Sivaji; FERACO, Antonio (Ed.). **A Practical Guide to Sentiment Analysis**. 5. ed. [s.l.]: Springer, 2017. Cap. 1. p. 1-10. (Socio-Affective Computing). Disponível em: <http://ww.sentic.net/practical-guide-to-sentiment-analysis.pdf>. Acesso em: 20 de set de. 2018.

CHRISTHIE, William. **SENTIMENTALL: FERRAMENTA PARA ANÁLISE DE SENTIMENTOS EM PORTUGUÊS**. 2015. 139 f. Trabalho de Conclusão de Curso (Graduação) – Curso de Sistemas de Informação, Centro Universitário Luterano de Palmas, Palmas/TO, 2015. Disponível em: <https://ulbra-to.br/bibliotecadigital/publico/home/documento/151>. Acesso em 20 de set de 2018.

COPPIN, Ben. **Inteligência Artificial**. Rio de Janeiro: Ltc, 2013. 636 p.

CORINNA, Cortes and Vladimir Vapnik. **Support-vector networks. Mach. Learn**., 20(3):273297, September 1995. 2. Disponsível em: <http://image.diku.dk/imagecanon/material/cortes_vapnik95.pdf>. Acesso em: 03 dez 2018.

CONGALTON, R. G. **A review of assessing the accuracy of classifications of remotely sensed data**. Remote Sensing of Environment, v. 49 n. 12, p. 1671-1678, 1991.

CANDELA, G., FIGINI, P. **The Economics of Tourism Destinations**, Springer, Heidelberg, 2012. Disponível em: <http://www.beck-shop.de/fachbuch/leseprobe/9783642208737_Excerpt_001.pdf>. Acesso em: 22 de set de 2018.

FIGUEIREDO, Geíza Coutinho; VIEIRA, Carlos Antonio Oliveira. **Estudo do comportamento dos índices de Exatidão Global, Kappa e Tau, comumente usados para avaliar a classificação de imagens do sensoriamento remoto**. Anais XIII Simpósio Brasileiro de Sensoriamento Remoto, Florianópolis, p. 5755-5762, abril 2007. Florianópolis, Brasil, 21-26 abril 2007. Disponível em: <http://marte.sid.inpe.br/col/dpi.inpe.br/sbsr@80/2006/11.13.17.35/doc/5755-5762.pdf>. Acesso em: 15 de maio de 2019.

FELDMAN, Ronen. **Techniques and Applications for Sentiment Analysis**. Communications Of The Acm, v. 56, n. 4, p.82-89, abr. 2013. Disponível em: <https://dl.acm.org/citation.cfm?id=2436274>. Acesso em: 27 set. 2018.

FOODY, G. M. **Status of land cover classification accuracy assessment**. Remote Sensing of Environment, v. 80, p. 185– 201, 2002.

HUSSEIN, Doaa Mohey El-din Mohamed. **A survey on sentiment analysis challenges.** Journal Of King Saud University: Engineering Sciences. Cairo, Egito, 9 p. abr. 2016. Disponível em: <https://www.researchgate.net/profile/Doaa_Mohey_El-Din/publication/301649355_A_Survey_on_Sentiment_Analysis_Challenges/links/571fa0d608aeaced788ac228/A-Survey-on-Sentiment-Analysis-Challenges.pdf?origin=publication_detail>. Acesso em 24 de set de 2018.

JÚNIOR, Severino Domingos da Silva; COSTA, Francisco José. **Mensuração e Escalas de Verificação: uma Análise Comparativa das Escalas de Likert e Phrase Completion**. PMKT - Revista Brasileira de Pesquisas de Marketing, Opinião e Mídia, São Paulo, p. 1-16, 24 jun. 2014.

LIU, B. (2010). **Sentiment analysis and subjectivity**. Disponível em: <https://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf>. Acesso em: 20 de set de 2018.

LOPES, L.; OLIVEIRA, L. H. M. DE; VIEIRA, R. **Extração de Métricas e Análise de Sentimentos em Comentários Web do Domínio de Hotéis**. 2010. Disponível em: <http://www.inf.ufg.br/sbsi2015/sites/portal.inf.ufg.br.sbsi2015/files/SBSI2015-Anais-Tracks-pag-001-259.pdf>. Acesso em: 29 nov 2018.

LIU, Bin. **Web Data Mining:** Exploring Hyperlinks, Contents, and Usage Data (DataCentric Systems and Applications). Springer, 2008. Disponível em: <https://downloadnema.com/wp-content/uploads/2017/03/web-data-mining-www.downloadnema.com_.pdf>. Acesso em: 03 dez 2018.

LEAL, Matheus R; CHRISTHI, Willian C.; BRITO, Parcilene F. **Avaliação de Desempenho de Uma Ferramenta de Análise de Sentimentos Baseada em Aspectos**. 2017. 12 f. ENCOINFO - 2017.

LIU, Bing. (2012). **Sentiment Analysis and Opinion Mining**. Morgan & Claypool Publishers. 167 p. Disponível em: <https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf>. Acesso em 13 de set de 2018.

LIRA, Sachiko Araki. **ANÁLISE DE CORRELAÇÃO: ABORDAGEM TEÓRICA E DE CONSTRUÇÃO DOS COEFICIENTES COM APLICAÇÕES**. 2004. 196 f. Dissertação apresentada ao Curso de Pós-Graduação. Universidade Federal do Paraná. Disponível em: <http://www.ipardes.gov.br/biblioteca/docs/dissertacao_sachiko.pdf>. Acesso em: Acesso em: 20 abr. 2019.

MARTINS, Nuno Miguel Nogueira. **Utilização de imagens de satélite de alta resolução para a extracção de elementos em ambiente urbano.** 89 f. Mestrado em Engenharia Geográfica. Departamento de Engenharia Geográfica, Geofísica e Energia Faculdade de Ciências da Universidade de Lisboa, 2012. Disponível em: <http://repositorio.ul.pt/bitstream/10451/9148/1/ulfc104672_tm_Nuno_Martins.pdf>. Acesso em: 7 de maio de 2019.

MCDONALD, Ryan; HANNAN, Kerry; NEYLON, Tyler; WELLS, Mike, REYNAR, Jeff (2007). **Structured Models for Fine-to-Coarse Sentiment Analysis** Disponível em: <http://www.aclweb.org/anthology/P07-1055>. Acesso em 16 de set de 2018.

MCCALLUM, Andrew, NIGAM, Kamal. **A comparison of event models for naive Bayes text classification**. In Learning for Text Categorization: Papers from the 1998 AAAI Workshop, volume 752, pages 4148, 1998. Disponível em: <http://www.cs.cmu.edu/~knigam/papers/multinomial-aaaiws98.pd>. Acesso em: 03 dez 2018.

NIGAM, K.; LAFFERTY, J.; MCCALLUM, A., **Using maximum entropy for text classification**, IJCAI-99 workshop on machine learning for information filtering, 2009. Disponível em: <https://www.cc.gatech.edu/~isbell/reading/papers/maxenttext.pdf>. Acesso em: 03 dez 2018.

NIVRE, Joakim. **Dependency Grammar and Dependency Parsing**. 2005. Disponível em: <http://stp.lingfil.uu.se/~nivre/docs/05133.pdf>. Acesso em: 21 abr. 2019.

OLIVEIRA, Taylor Santos. **Implementação de um módulo de inferência da polaridade geral dos comentários do TripAdvisor analisados na SentimentALL.** 2018. 25f. Estágio Supervisionado em Ciência da Computação. (Bacharel em Ciência da Computação) - Ceulp Ulbra, Palmas.

OLIVEIRA, Graziela da Silva Rocha. **Avaliação da Qualidade de Resultados Obtidos Através dos Métodos de Classificação Supervisionada - Máxima Verossimilhança e Redes Neurais**. Curso de Pós-Graduação em Geoprocessamento, Departamento de Cartografia, Instituto de Geociências. Universidade Federal de Minas Gerais. Belo Horizonte - 2003. 35 f. Disponível em: <http://www.csr.ufmg.br/geoprocessamento/publicacoes/graziela2003.pdf>. Acesso em: 14 de maio de 2019.

SILVA, Nelson Rocha; LIMA, Diego; BARROS, Flávia. **SAPair: Um Processo de Análise de Sentimento no Nível de Característica**. In: BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEM, 2012, Disponível em: <http://sites.labic.icmc.usp.br/wti2012/artigos/105283.pdf>. Acesso em: 28 nov 2018.

SANTOS, Leandro Matioli. **Protótipo para mineração de opiniões em redes sociais: estudo de casos selecionados usando o twitter**. Trabalho de Graduação - Universidade Federal de Lavras - MG, 2010. Disponível em: <http://repositorio.ufla.br/bitstream/1/5190/1/MONOGRAFIA_Prototipo_para_mineracao_de_opiniao_em_redes_sociais_estudo_de_casos_selecionados_usando_o_twitter.pdf>. Acesso em: 03 dez 2018.

SILVA, Lucas Lo Ami Alvino. **Análise de sentimentos em contexto: estudo de caso em blog de empreendedorismo**. 67 f. Monografia - Universidade de Brasília - DF, 2013. Disponível em:

<http://bdm.unb.br/bitstream/10483/10153/1/2013_LucasLoAmiAlvinoSilva.pd> . Acesso em: 03 dez 2018.

KAUER, Anderson Uilian. **Análise de Sentimentos baseada em Aspectos e Atribuição de Polaridade.** 2016. 76 f. Dissertação (Mestrado em Ciência da Computação) - INSTITUTO DE INFORMÁTICA PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO, UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL, Porto Alegre, 2016. Disponível em: <https://www.lume.ufrgs.br/bitstream/handle/10183/140910/000991520.pdf?sequence=1>. Acesso em : 20 de set de 2018.

KALAIVANI, P.; SHUNMUGANATHAN, K. L. **SENTIMENT CLASSIFICATION OF MOVIE REVIEWS BY SUPERVISED MACHINE LEARNING APPROACHES**. Indian Journal Of Computer Science And Engineering. Indian, p. 285-292. set. 2014. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=192A412E567C2E1A39598674056B569C?doi=10.1.1.681.6806&rep=rep1&type=pdf>.Acesso em: 28 nov 2018.

KIM, S.-M.; HOVY, E. **Determining the sentiment of opinions**. In Proceedings of the International Conference on Computational Linguistics (COLING 2004). 7 f. East Stroudsburg, PA: Association for Computational Linguistics. Disponível em: <https://dl.acm.org/citation.cfm?id=1220555>. Acesso em: 11 de maio de 2019.

WANG, Dong; LIU, Yang. **A cross-corpus study of unsupervised subjectivity identification based on calibrated EM.** In: Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, 2., 2011, Portland. Proceedings. Stroudsburg, Pa: Association For Computational Linguistics, 2011. p. 161 - 167. Disponível em: <https://dl.acm.org/citation.cfm?id=2107674>. Acesso em: 29 nov 2018.

ZHANG, L.; WANG, S.; LIU, B. **Deep learning for sentiment analysis: A survey.** Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, p. e1253, 2018. Disponível em: <https://arxiv.org/ftp/arxiv/papers/1801/1801.07883.pdf> . Acesso em: 28 nov 2018.

ZHANG, Changli et al**. Sentiment Analysis of Chinese Documents: From Sentence to Document Level.** 2011. 15 f. Journal of the American Society for Information Science and Technology. Disponível em: <https://www.researchgate.net/publication/220433795Ver>. Acesso em: 22 de set de 2018.