Estudo de Técnicas de Análise de Sentimento em *Reviews* de Hotelaria

Alice Silva de Souza, Márcio de Souza Dias

Resumo Reserva de Hotéis pela internet tem se tornado comum no Brasil, e a decisão de compra dos consumidores passaram a ser influenciadas pelos comentários deixados pelos usuários na web. A grande quantidade de comentários gerados na internet tem tornado difícil a avaliação manual dos comentários sobre o produto, se são comentários positivos ou negativos. Diante disso, esse trabalho propõe a análise de sentimento de comentários na web no âmbito de hotelaria de forma automatizada. Utilizando um corpus de comentários de hotéis, o trabalho obteve uma acurácia de 83,11%.

1 Introdução

A análise de sentimento, é o campo de estudo em computação aplicada em processamento de linguagem natural (PLN), que procura extrair, avaliações, opiniões, emoções e comportamentos associados a serviços, produtos e assuntos apresentados em um texto, entre outros termos (MACHADO, 2018). Na literatura Liu (2012) define a análise de sentimentos como: ”área direcionada para a análise de opiniões, sentimentos, avaliações, atitudes e emoções das pessoas da linguagem escrita”.

Atualmente o Brasil é o maior mercado de internet da América Latina e o quarto maior mercado de internet do mundo em número de usuários de internet (STATISTA, 2020). O número de sites de *e-commerce* no Brasil cresceu mais de 37% no último ano (2018-2019), chegando a cerca de 930 mil sites de comércio virtual (PAYPAL, 2019). Segundo Cortez e Mondo (2017), o processo de decisão de compra dos consumidores passou a ser tão, ou mais, influenciado pelas opiniões dos internautas, do que as informações institucionais publicadas na *web*. Devido ao aumento das compras *online* e a grande quantidade de comentários gerados tornou-se difícil a identificação manual do sentimento expressado nos comentários sobre o produto, assim, fez-se essencial processar as opiniões disponíveis na *web* de forma automatizada.

De acordo com Kasper e Vela (2011), planejamento de viagens e reserva de hotéis na *web* tem se tornado cada vez mais comuns e rotineiros. Especialmente se tratando de reservas de hotéis, os comentários deixados pelos usuários são extremamente relevantes, uma vez que são

Alice Silva de Souza
Instituto de Biotecnologia, Universidade Federal de Catalão, Catalão, Goiás, Brasil.
e-mail: alicesouza@discente.ufg.br

Márcio de Souza Dias
Instituto de Biotecnologia, Universidade Federal de Catalão, Catalão, Goiás, Brasil.
e-mail: marciosouzadias@ufg.br

Anais do XV Encontro Anual de Ciência da Computação (EnAComp 2020). ISSN: 2178-6992.
Catalão, Goiás, Brasil. 25 a 27 de Novembro de 2020.

mais sinceros e não são baseados em *marketing*, como a página online, ou um catálogo do hotel, já que o comentário contém a real opinião de um cliente sobre o hotel. Portanto, a análise de sentimentos no ramo de hotelaria é importante tanto para o usuário conhecer da qualidade do hotel em que ficará hospedado, quanto para o próprio gerenciamento do hotel.

O projeto tem como objetivo processar de forma automatizada comentários disponíveis na *web* para textos em português no âmbito de hotelaria, utilizando técnicas de análise de sentimento que buscam identificar o sentimento contido no *review*. Além disso, também tem como propósito avaliar e comparar as ferramentas utilizadas no trabalho, afim de, no futuro criar uma ferramenta eficiente de avaliação de opiniões de comentários de hotéis, para auxiliar na escolha do consumidor sobre determinado hotel o qual deseja se hospedar e o hotel ter um retorno sobre as avaliações que os hospedes tiveram sobre o hotel.

Este artigo está organizado da seguinte forma: na Seção 2 abordamos os trabalhos relacionados. Na Seção 3 apresentamos o corpus e os recursos utilizados. Na Seção 4 o desenvolvimento do trabalho. Na Seção 5 os experimentos e as análises. Por último, na Seção 6 abordamos a conclusão do trabalho e trabalhos futuros.

2 Trabalhos Relacionados

O trabalho de Taboada et al. (2011) utiliza de métodos baseado em léxicos para identificação de sentimento e mostra uma maneira de cálculo de orientação semântica (*Semantic Orientation Calculator* - SO-CAL). O método é feito calculando a polaridade de um texto usando um dicionário de sentimento e, além disso, reconhece palavras denominadas de modificadores, que intensificam, invertem ou neutralizam essas polaridades.

O trabalho de Freitas et al. (2015) propõe uma técnica de análise de sentimento a nível de aspecto para comentários em português brasileiro no setor de hotelaria. Com o intuito de alcançar uma avaliação completa, os autores reconhecem aspectos explícitos e implícitos utilizando ontologias. A técnica é realizada em quatro etapas: pré-processamento, identificação de aspecto, identificação de polaridade e sumarização.

Outro importante trabalho foi o de Avanço (2015) que desenvolveu técnicas e sistemas que buscam a normalização de comentários na *web* (tratamento do texto com modificação de “internetês” e correção ortográfica e de pontuação), e a categorização de *reviews*, no âmbito de produtos, a nível de texto para o português brasileiro. Para a normalização, o autor usou de procedimentos linguísticos e, para a categorização de *reviews*, atribuiu a polaridade positiva ou negativa aos comentários usando léxicos e aprendizagem de máquina, tal como a junção de ambas na elaboração de um recurso híbrido original. Com isso, os autores comprovaram que se obtêm os melhores resultados com a normalização para técnicas baseadas em léxico.

3 Corpus e Recursos

Representante da Linguística de Corpus, Sardinha (2004) adota a definição proposta por Sánchez e Cantos (1996), que assim definem corpus: “um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados

critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos na totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados úteis para a descrição e análise." (p. 8-9).

O corpus já normalizado utilizado neste trabalho é o corpus TripAdvisor (SOUZA; OLIVEIRA; ALEXANDRA, 2018), o qual foi extraído do site TripAdvisor e consiste em avaliações de hotéis das capitais dos estados brasileiros e do Distrito Federal que datam de 27/05/2004 até 20/03/2018. Tal corpus contém 730.069 *reviews*, 55.950.007 tokens e 457.337 tipos.

Souza et al. (2018) também utilizou o corpus de Souza e Vieira (2011) e observou que cada *review*, em média, contém aproximadamente 77 tokens, com o maior *review* possuindo 2.857 tokens e a menor possuindo apenas dois. A avaliação de cada *review* foi dada pela quantidade de estrelas: 1 e 2 estrelas, a polaridade do comentário é atribuído como negativo; 3 estrelas como neutro; e 4 e 5 estrelas como positivo. Souza et al. (2018) observou, que a classe negativa corresponde a 7,1% do corpus, 16% das avaliações são neutras e 76,2% são positivas.

Para este trabalho, utilizamos três léxico de sentimentos o Sentilex (SILVA; CARVALHO; SARMENTO, 2012), o LIWC (BALAGE FILHO; PARDO; ALUÍSIO, 2013) e o OpLexicon (SOUZA; VIEIRA, 2011). Segundo Avanço (2015), léxicos de sentimentos são compostos por um agrupamento de palavras de uma língua geralmente usada para expressar sentimento.

O SentiLex (SILVA; CARVALHO; SARMENTO, 2012) é um léxico de sentimento criado para a análise de sentimento e opinião sobre entidades humanas em textos redigidos em português. Atualmente a ferramenta é formada por 7.014 lemas e 82.347 formas flexionadas. O léxico SentiLex (SILVA; CARVALHO; SARMENTO, 2012) contém: 16.863 adjetivos; 1.280 substantivos, 29.504 verbos e 34.700 expressões idiomáticas.

O dicionário Brazilian Portuguese LIWC (BALAGE FILHO; PARDO; ALUÍSIO, 2013) foi construído a partir do dicionário em inglês LIWC (PENNEBAKER; FRANCIS; BOOTH, 2001). O LIWC (BALAGE FILHO; PARDO; ALUÍSIO, 2013) contém 127.149 entradas que não são morfologicamente categorizadas e a classificação das entradas é diferente: das 127.149 entradas, 12.878 são positivas e 15.115 são negativas, e além disso, encontra-se categorias como: palavras de negação (19), saúde (7.003), amizade (679), dinheiro (5.353), família (96), amizade (679), entre outras, totalizando 64 categorias.

Outro léxico utilizado, o OpLexicon (SOUZA; VIEIRA, 2011) possui 30.322 palavras (23.433 adjetivos e 6.889 verbos) e foi criado a partir de diferentes técnicas. O primeiro léxico foi desenvolvido usando a avaliação de um corpus anotado. O outro foi construído usando pesquisa de antônimos e sinônimos através de uma lista inicial de palavras. O último léxico foi obtido por meio da tradução automática do Liu's English Opnion Lexicon (HU; LIU, 2004). Os resultados de cada uma dessas técnicas foram combinados para gerar um grande léxico para o português do Brasil o Oplexicon (SOUZA; VIEIRA, 2011).

Além dos léxicos, também utilizamos como ferramenta etiquetadores morfossintáticos para o português, que são o: TreeTagger[1] e o spaCy[2]. De acordo com Domingues, Favero e Medeiros (2008), etiquetagem morfossintática é o trabalho de detectar as categorias gramaticais das palavras em uma sentença.

Para este trabalho foi utilizado o Classificador de Opiniões baseado em Léxico (CBL) (AVANÇO, 2015), que a priori identifica a polaridade das palavras de sentimento estabelecida pelo léxico de

[1] https://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

[2] https://spacy.io/

sentimento. Se na vizinhança dessas palavras ocorrerem palavras de negação (e.g. “jamais”, “nada”, “nem”, “nenhum”, “ninguém”, “nunca”, “não”, “tampouco”), intensificação (e.g. “mais”, “muito”, “demais”, “completamente”, “absolutamente”, “totalmente”, “definitivamente”, “extremamente”, “frequentemente”, “bastante”) ou redução (e.g. “pouco”, “quase”, “menos”, “apenas”), as respectivas polaridades são modificadas. Logo depois, é calculada a orientação semântica do texto ou da sentença, através da adição das polaridades encontradas, possivelmente alteradas pelo contexto.

No algoritmo de CBL (AVANÇO, 2015), uma palavra de sentimento terá sua polaridade invertida quando a mesma estiver presente em um contexto em que apenas ocorra negação. Quando estiver em um contexto onde há uma palavra de intensificação sem uma palavra de negação, a polaridade da palavra é triplicada; se tiver intensificação junto com negação a polaridade é dividida por três. E quando estiver em um contexto que há uma palavra de redução sem uma palavra de negação, a polaridade é dividida por três; se tiver redução junto com negação a polaridade é triplicada. O fator de intensificação e redução igual a três foi determinado por experimentos. A Figura 1 demonstra um exemplo de um contexto em que ocorre um caso de intensificação e negação em uma sentença.

”Hotel perfeito para apreciar a praia, mas o café não é muito bom.”
perfeito(+1) + apreciar(+1) + não é muito bom(1/3) = +2.33

Figura 1: Cálculo da polaridade no algoritmo CBL.

Na Figura 1, há três palavras de sentimento: *perfeito*, *apreciar* e *bom*. A primeira e a segunda não são afetados pela palavra de negação, intensificação e redução, mantendo a sua polaridade. Já a terceira sofre influência da palavra de negação (não) e de intensificação (muito) reduzindo a força da sua polaridade.

Outro método que utilizamos, foi o de análise de sentimento de Machado (2018), que faz uso dos léxicos: Sentilex (SILVA; CARVALHO; SARMENTO, 2012), OpLexicon (SOUZA; VIEIRA, 2011), LIWC (BALAGE FILHO; PARDO; ALUÍSIO, 2013) e LexReli (criado por Machado (2018)). O autor utiliza dos léxicos para análise de sentimento, primeiro de forma independente uns dos outros; depois na combinação dos três primeiros léxicos citados em ordens diferentes; e por último, acrescenta no começo dos léxicos combinados o léxico LexReli (MACHADO, 2018), com o objetivo de avaliar qual das ferramentas terá o melhor resultado.

Para a avaliação do *review*, Machado (2018) utiliza e analisa várias técnicas. Uma delas é o método da polaridade das palavras, que basicamente é a busca das palavras do texto no léxico de sentimento, independentemente de sua classe gramatical. Para se obter o resultado, soma a polaridade das palavras: se for maior que zero, o comentário é positivo, e se for menor que zero, o comentário é negativo, como mostra a Figura 2.

Na Figura 2, há três palavras de sentimento: *perfeito*, *apreciar* e *bom*. A primeira e a segunda não são afetados pela palavra de negação, mantendo a sua polaridade. Já a terceira sofre influência da palavra de negação (não), invertendo sua polaridade.

Outro método utilizado é o “somente adjetivos”, parecido com a técnica anterior, mas, como o próprio nome já diz, soma a polaridade somente dos adjetivos, como mostra a Figura 3.

”Hotel perfeito para apreciar a praia, mas o café não é muito bom.”
perfeito(+1) + apreciar(+1) + (não(-1) * bom(+1)) = 1

Figura 2: Cálculo da polaridade, utilizando o método polaridade das palavras.

”Hotel perfeito para apreciar a praia, café da manhã maravilhoso.”
perfeito(+1) + maravilhoso(+1) = 2

Figura 3: Cálculo da polaridade, utilizando o método somente adjetivos.

Na Figura 3, apenas identifica-se as palavras de sentimento *perfeito* e *maravilhoso*, pois são palavras adjetivas, não detectando a palavra de sentimento *apreciar*, por ser um verbo. A primeira e a segunda palavra identificadas, não sofrem influência de nenhuma palavra de negação, mantendo sua polaridade.

Por último, outra técnica utilizada é a de preferência aos adjetivos, primeiro é considerado as polaridades dos adjetivos na frase, e se não houver adjetivo o algoritmo verifica todas as palavras da frase, como mostram as Figuras 4 e 5.

Hotel para descansar e relaxar.
descansar(+1) + relaxar(+1) = 2

Figura 4: Cálculo da polaridade, utilizando o método preferência aos adjetivos.

Na Figura 4, o primeiro passo é identificar as palavras de sentimento que são adjetivos. Porém, como não há palavras de sentimento que são adjetivos na sentença, a técnica passa a procurar todas as palavras que são palavras de sentimento, independentemente da sua classe gramatical, detectando assim, as palavras de sentimento *descansar* e *relaxar*. Por não conter negação no contexto a primeira e a segunda palavra detectada não tem sua polaridade invertida.

”Hotel perfeito para apreciar a praia, café da manhã maravilhoso.”
perfeito(+1) + maravilhoso(+1) = 2

Figura 5: Cálculo da polaridade, utilizando o método preferência aos adjetivos.

Na Figura 5, que utiliza a mesma técnica da Figura 4, apenas identifica-se as palavras de sentimento *perfeito* e *maravilhoso*, pois são palavras adjetivas, não precisando passar para a etapa de identificar outras palavras de sentimento que não são adjetivos, como ocorre na Figura 4. Assim, a palavra de sentimento *apreciar* não é identificada, por ser um verbo. A primeira e a segunda palavra identificadas não sofrem influência de nenhuma palavra de negação, mantendo sua polaridade.

Em seguida, o autor utiliza de todos esses métodos junto com aspectos. Primeiramente ele identifica o aspecto da sentença e em seguida a palavra de sentimento mais próximo a ele, atribuindo esse sentimento ao aspecto. Em todos os métodos, Machado (2018) verifica se há alguma palavra de negação anterior a palavra de sentimento, e caso exista, a polaridade da palavra é invertida.

4 Desenvolvimento do trabalho

Para o desenvolvimento do trabalho realizamos 2 etapas: pré-processamento e identificação de sentimento.

4.1 Pré-Processamento

Pré-processamos do corpus TripAdvisor (SOUZA; OLIVEIRA; ALEXANDRA, 2018), separamos o corpus em positivo, negativo e neutro baseado em suas estrelas, conforme o trabalho de Souza et al. (2018): 1 e 2 estrelas (negativo); 4 e 5 estrelas (positivo) e 3 estrelas (neutros). Os neutros descartamos, gerando um corpus de 608.834 *reviews*. Com isso, observamos que o corpus estava consideravelmente desbalanceado pois 91,45% dos *reviews* eram positivos e 8,54% eram negativos. Para o equilíbrio, de forma aleatória, a quantidade de *reviews* positivo foi reduzida, equiparando-se à quantidade de *reviews* negativo, criando um corpus de 104.000 comentários. O corpus resultante foi separado em 70% para treino e 30% para teste.

Além do pré-processamento do corpus, também, pré-processamos os léxicos de sentimentos: Sentilex (SILVA; CARVALHO; SARMENTO, 2012), OpLexicon (SOUZA; VIEIRA, 2011) e LIWC (BALAGE FILHO; PARDO; ALUÍSIO, 2013). O primeiro passo desse pré-processamento é a retirada das acentuações das palavras de sentimento, pois em alguns casos o *review* contém alguns erros de acentuação, então, para que todas as palavras de sentimento sejam identificadas, durante a detecção de sentimento, também retiramos a acentuação das palavras do *review* que contenham acentuação. Em seguida, combinamos todos os léxicos de sentimentos em diversas ordens diferentes, com o objetivo de utilizar os métodos de análise de sentimento de Machado (2018).

4.2 Identificação de Sentimento

Para identificar o sentimento, primeiramente utilizamos algum *tagger* para marcar a classe gramatical de cada palavra do *review*, e pré-processamos cada palavra retirando suas acentuações e transformamos as letras maiúsculas em minúsculas. Em seguida, utilizamos os métodos de Machado (2018), que são: polaridade das palavras, somente adjetivos e preferência aos adjetivos. Em todos os métodos verificamos se havia alguma palavra de negação anterior a palavra de sentimento, e caso houvesse, a polaridade da palavra era invertida.

Além disso, utilizamos a ferramenta CBL de Avanço (2015), que faz o uso de léxicos separadamente, mas acrescentamos a ideia de Machado (2018) de utilizar vários dicionários de sentimentos juntos e as técnicas: polaridade das palavras, somente adjetivos e preferência aos adjetivos.

Para cada técnica, testamos os *taggers* spaCy e TreeTagger. Usamos os léxicos de sentimentos separadamente e também combinando-os em diferentes ordens assim como no trabalho de Machado (2018). Verificamos tamanhos de janelas diferentes para identificação de palavras de negação para a técnica de Machado (2018) e também analisamos tamanho de janelas diferentes para identificação de palavras de negação, redução e intensificação para a técnica de Avanço (2015).

Para analisar quais são as melhores técnicas, baseamos em cálculos utilizados por Avanço (2015), como mostra na Figura 6, as quais são: Precisão para a classe positiva (1), Precisão para a classe negativa (2), Cobertura para classe positiva (3), Cobertura para classe negativa (4), F1-Média (5) e Acurácia (6). Os valores podem ser adquiridos através da matriz de confusão (Tabela 1), que é basicamente uma tabela que determina os tipos de acertos e erros realizados em um trabalho de classificação (tradicionalmente binária). Os valores são melhores quanto mais próximas de 1 e piores quanto mais próximas de 0. Na Tabela 1, “VP”, “FP”, “FN” e “VN” significam respectivamente: verdadeiro positivo, falso positivo, falso negativo e verdadeiro negativo.

		Humano	
		Positivo	Negativo
Máquina	Positivo	VP	FP
	Negativo	FN	VN

Tabela 1: Matriz de confusão.

O cálculo das medidas utilizadas para avaliação da classificação de sentimento foi realizado individualmente para cada classe (positivos, negativos).

$$(1)\ P+ = \frac{VP}{VP+FP} \quad (2)\ P- = \frac{VN}{VN+FN} \quad (3)\ C+ = \frac{VP}{VP+FN} \quad (4)\ C- = \frac{VN}{VN+FP}$$

$$(5)\ F1 = 2.\frac{P.C}{P+C} \quad (6)\ A = \frac{VP+VN}{VP+FN+FP+VN}$$

Figura 6: Cálculo da polaridade, utilizando o método preferência aos adjetivos.

5 Experimentos e Análises

Neste trabalho, para detecção de palavras de negação, intensificação e redução no *review*, foi utilizado como contexto uma janela de quatro palavras. Se palavra negação, intensificação e redução estiver até três palavras atrás de uma palavra de sentimento a polaridade da palavra

de sentimento é modificada. Definimos o tamanho de janela quatro por meio de experimentos. Fizemos testes com janelas de três ou cinco palavras e obtivemos resultados inferiores.

Na Tabela 2, mostramos os resultados dos experimentos realizados com o *tagger* TreeTagger. Já na Tabela 3, mostramos os resultados dos experimentos realizados com o *tagger* Spacy. Na primeira parte de ambas as tabelas, apresentamos os resultados dos léxicos sozinhos, e na segunda parte das tabelas apresentamos resultados para a combinação dos léxicos. Nas duas partes, apresenta-se resultados junto com as técnicas de: polaridade das palavras, somente adjetivos e preferência aos adjetivos, tanto utilizando palavras de negação como modificador da polaridade das palavras de sentimento, como utilizando palavras de negação, intensificação e redução como modificadores da polaridade das palavras de sentimento.

Observa-se que na primeira parte da Tabela 2, o melhor resultado obtido foi da combinação das palavras de negação como modificadores da polaridade das palavras de sentimento, com a técnica polaridade das palavras, juntamente com o léxico SentiLex (SILVA; CARVALHO; SARMENTO, 2012), resultando em uma acurácia de 82,95%. Já na segunda parte da tabela o melhor resultado obtido foi da seguinte combinação: os léxicos (Sentilex (SILVA; CARVALHO; SARMENTO, 2012), OpLexicon (SOUZA; VIEIRA, 2011) e LIWC (BALAGE FILHO; PARDO; ALUÍSIO, 2013)), juntamente com a técnica polaridade das palavras e palavras de negação, intensificação e redução como modificadores da polaridade das palavras de sentimento, resultando em uma acurácia de 76,89%.

Observa-se que, mesmo o melhor resultado sendo na primeira parte da tabela usando palavras de negação como modificadores da palavra de sentimento, os léxicos OpLexicon (SOUZA; VIEIRA, 2011) e LIWC (BALAGE FILHO; PARDO; ALUÍSIO, 2013) obtiveram melhores resultados na segunda parte da tabela, combinados entre si (e com o léxico SentiLex (SILVA; CARVALHO; SARMENTO, 2012)) e utilizando palavras de negação, intensificação e redução como modificadores da palavra de sentimento.

As Tabelas 2 e 3 utilizam de algumas abreviações como: F1-Média (F1); Acurácia (Ac); Polaridade das Palavras (Pal); Somente Adjetivos (Adj); Preferência aos Adjetivos (P.Adj); Intensificação (Int).

	TreeTagger - Negação						TreeTagger - Negação, Int. e Redução					
Técnicas	Pal	Pal	Adj	Adj	P.Adj	P.Adj	Pal	Pal	Adj	Adj	P.Adj	P.Adj
Léxico	F1	Ac	F1	Ac	F1	Ac	F1	Ac	F1	Ac	F1	Ac
OpLexicon	72,69	74,07	71,19	73,07	71,37	73,21	73,55	74,78	72,24	73,88	72,41	74,02
SentiLex	**82,74**	**82,95**	**75,33**	**76,48**	**75,54**	**76,65**	**82,64**	**82,85**	**76,05**	**77,06**	**76,26**	**77,23**
LIWC	60,85	65,07	59,79	64,31	59,97	64,43	62,3	66,06	61,18	65,24	61,36	65,36
Combinação												
LIWC+Op+Sent	71,19	72,89	68,93	71,26	69,14	71,42	71,74	73,35	69,77	71,88	69,98	72,05
LIWC+Sent+Op	72,96	74,32	71,33	73,16	71,53	73,32	73,78	75,03	72,30	73,91	72,51	74,07
Op+LIWC+Sent	73,04	74,38	71,19	73,07	71,41	73,24	73,90	75,13	72,23	73,87	72,45	74,05
Op+Sent+LIWC	72,84	74,20	71,19	73,07	71,40	73,23	73,70	74,94	72,23	73,87	72,44	74,04
Sent+LIWC+Op	73,77	74,96	73,10	74,60	73,20	74,75	74,65	75,71	74,00	75,30	74,19	75,46
Sent+Op+LIWC	**74,86**	**75,87**	**74,44**	**75,73**	**74,63**	**75,88**	**76,02**	**76,89**	**75,59**	**76,66**	**75,78**	**76,82**

Tabela 2: Resultados da análise de sentimento usando o léxico TreeTagger.

Na Tabela 3, o melhor resultado obtido foi da combinação do léxico SentiLex (SILVA; CARVALHO; SARMENTO, 2012), com a técnica polaridade das palavras e das palavras de negação como modificador da polaridade das palavras de sentimento, resultando em uma acurácia de

83,11%. Já na segunda parte da tabela o melhor resultado obtido foi da combinação dos léxicos (Sentilex (SILVA; CARVALHO; SARMENTO, 2012), OpLexicon (SOUZA; VIEIRA, 2011) e LIWC (SOUZA; VIEIRA, 2011)), da técnica polaridade das palavras e palavras de negação, intensificação e redução como modificadores da polaridade das palavras de sentimento, resultando em uma acurácia de 77,03%.

Nota-se que assim como ocorreu na Tabela 2, o melhor resultado na Tabela 3 foi na primeira parte da tabela usando palavras de negação como modificadores da polaridade das palavras de sentimento e o léxico SentiLex (SILVA; CARVALHO; SARMENTO, 2012). Observa-se também que os léxicos OpLexicon (SOUZA; VIEIRA, 2011) e LIWC (SOUZA; VIEIRA, 2011) obtiveram melhores resultados na segunda parte da tabela combinados entre si (e com o léxico SentiLex (SILVA; CARVALHO; SARMENTO, 2012)), e maior acurácia utilizando palavras de negação, intensificação e redução como modificadores da palavra de sentimento em ambas as partes da tabela.

	Spacy - Negação						Spacy - Negação, Int. e Redução					
Técnicas	Pal	Pal	Adj	Adj	P.Adj	P.Adj	Pal	Pal	Adj	Adj	P.Adj	P.Adj
Léxico	F1	Ac	F1	Ac	F1	Ac	F1	Ac	F1	Ac	F1	Ac
OpLexicon	72,90	74,25	68,99	71,25	69,32	71,50	73,75	74,96	69,93	71,94	70,28	72,21
SentiLex	**82,91**	**83,11**	**72,66**	**74,18**	**73,05**	**74,50**	**82,28**	**82,99**	**73,38**	**74,73**	**73,77**	**75,05**
LIWC	60,95	65,14	58,26	63,18	58,64	63,43	62,39	66,14	59,45	63,94	59,82	64,19
Combinação												
LIWC+Op+Sent	71,39	73,05	66,95	69,66	67,33	69,95	71,93	73,52	67,52	70,05	67,91	70,34
LIWC+Sent+Op	73,16	74,50	69,13	71,34	69,51	71,63	73,99	75,21	69,88	71,87	70,26	72,16
Op+LIWC+Sent	73,24	74,55	69,17	71,39	69,55	71,69	74,07	75,27	70,09	72,07	70,48	72,38
Op+Sent+LIWC	73,02	74,02	69,18	71,41	69,56	71,70	73,90	75,11	70,09	72,08	70,47	72,37
Sent+LIWC+Op	73,95	75,14	70,84	72,68	71,20	72,96	74,85	75,89	71,54	73,19	71,91	73,48
Sent+Op+LIWC	**75,07**	**76,05**	**72,31**	**73,89**	**72,67**	**74,18**	**76,18**	**77,03**	**73,26**	**74,61**	**73,62**	**74,91**

Tabela 3: Resultados da análise de sentimento usando o léxico Spacy.

Entre as duas tabelas, o método empregado na Tabela 3 foi o que obteve os melhores resultados, e o melhor resultado da Tabela 3 foi de 83,11% comparado ao melhor resultado da Tabela 2 que foi de 82,95%.

6 Conclusão

Objetivo deste projeto é implementar técnicas de análise de sentimento no âmbito de hotelaria para textos em português e analisar as ferramentas utilizadas. O projeto está em fase inicial, mas até o momento a maior acurácia obtida foi de 83,11%, utilizando o método de polaridade das palavras, o léxico Spacy e o *tagger* TreeTagger. Para os próximos passos, pretendemos melhorar as técnicas aplicadas e ir em busca de novos métodos como, utilização de aspectos e aprendizado de máquina para análise de sentimento, com o propósito de comparar ou incrementar as técnicas já implementados.

Referências

AVANÇO, Lucas Vinicius. **Sobre normalizaçao e classificaçao de polaridade de textos opinativos na web**. 2015. Tese (Doutorado) – Universidade de São Paulo.

BALAGE FILHO, Pedro; PARDO, Thiago Alexandre Salgueiro; ALUÍSIO, Sandra. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In: PROCEEDINGS of the 9th Brazilian Symposium in Information and Human Language Technology. [S.l.: s.n.], 2013.

CORTEZ, Maria Carolina Antunes; MONDO, Tiago Savi. Comentários On-line: Formação de Expectativa e Decisão de Compra de Consumidores Hoteleiros/Online Reviews: Formation of Expectation and Decision to Purchase of Hotel Consumers. **ROSA DOS VENTOS-Turismo e Hospitalidade**, v. 10, n. 1, 2017.

DOMINGUES, Miriam Lúcia; FAVERO, Eloi Luiz; MEDEIROS, Ivo Paixão de. O desenvolvimento de um etiquetador morfossintático com alta acurácia para o português. **Avanços da Linguística de Corpus no Brasil. São Paulo: Humanitas**, p. 267–286, 2008.

FREITAS, Larissa Astrogildo de et al. Feature-level sentiment analysis applied to brazilian portuguese reviews. Pontifícia Universidade Católica do Rio Grande do Sul, 2015.

HU, Minqing; LIU, Bing. Mining opinion features in customer reviews. In: 4. AAAI. [S.l.: s.n.], 2004. v. 4, p. 755–760.

KASPER, Walter; VELA, Mihaela. Sentiment analysis for hotel reviews. In: COMPUTATIONAL linguistics-applications conference. [S.l.: s.n.], 2011. v. 231527, p. 45–52.

MACHADO, Mateus Tarcinalli. **Estudo e avaliação de métodos de análise de sentimentos baseada em aspectos para textos opinativos em português**. 2018. Tese (Doutorado) – Universidade de São Paulo.

PAYPAL, Blog. **Pesquisa: e-commerce brasileiro cresceu 37,5% em um ano**. 2019. Disponível em: ¡https://www.paypal.com/stories/br/pesquisa-e-commerce-brasileiro-cresceu-37-5-em-um-ano¿.

PENNEBAKER, JW; FRANCIS, ME; BOOTH, RJ. **Linguistic Inquiry and Word Count. Mahwah, NJ: LEA Software and Alternative Media**. [S.l.]: Inc, 2001.

SÁNCHEZ, Aquilino; CANTOS, Pascual. **Cumbre - Curso de Espãnol. Madri**. [S.l.]: Sociedad general espanõla de librería, 1996.

SARDINHA, Tony Berber. **Linguística de corpus**. [S.l.]: Editora Manole Ltda, 2004.

SILVA, Mário J; CARVALHO, Paula; SARMENTO, Luís. Building a sentiment lexicon for social judgement mining. In: SPRINGER. INTERNATIONAL Conference on Computational Processing of the Portuguese Language. [S.l.: s.n.], 2012. p. 218–228.

SOUZA, Joana; OLIVEIRA, Alcione; ALEXANDRA, Moreira. Development of a Brazilian Portuguese Hotel's Reviews Corpus: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings. In: [s.l.: s.n.], jan. 2018. p. 353–361. ISBN 978-3-319-99721-6. DOI: 10.1007/978-3-319-99722-3_36.

SOUZA, Joana Gabriela Ribeiro de et al. Análise de sentimento por meio de aprendizado profundo aplicado a avaliações de hotéis. Universidade Federal de Viçosa, 2018.

SOUZA, Marlo; VIEIRA, Renata. Construction of a portuguese opinion lexicon from multiple resources. **Anais do Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, 2011, Brasil.**, 2011.

STATISTA. ***Internet usage in Brazil - Statistics e Facts.*** **Disponível em: https://www.statista.com/topics/2045/internet-usage-in-**

brazil/#dossierSummary__chapter4. Acesso em: 31/06/2020. 2020. Disponível em: ¡https://www.statista.com/topics/2045/internet-usage-in-brazil/%5C#dossierSummary__chapter4¿.

TABOADA, Maite et al. Lexicon-based methods for sentiment analysis. **Computational linguistics**, MIT Press, v. 37, n. 2, p. 267–307, 2011.