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

Yago A. Santos, Márcio de Souza Dias

Resumo Com o aumento das possibilidades utilizando um smartphone, computador ou tablet que disponha de internet, as formas de compra e venda atuais são diferentes como as que eram utilizadas décadas atrás. O consumidor tem cada vez mais facilidade de adquirir algo pela internet, e está mais preocupado com a qualidade do produto que está sendo adquirido. O trabalho aqui proposto tem como objetivo realizar um desenvolvimento pautado na avaliação de técnicas para análise de sentimento e que serão descritas neste artigo.

1 Introdução

O grande volume de dados no qual a internet vem acumulando ao longo das últimas décadas, desenha um bom cenário quando se trata de pesquisas. De acordo com Liu (2012), há um elevado número de dados nas mídias sociais na web, sendo que sem esses dados, muitas pesquisas não teriam sido possíveis. Estando a análise de sentimentos no centro da pesquisa de mídias sociais, esses avanços não causam impactos apenas em PLN - Processamento de Linguagem Natural, mas também em diversas áreas da ciência.

Segundo Gonzalez, PLN trata computacionalmente os diversos aspectos da comunicação humana, como sons, palavras, sentenças e discursos, considerando formatos e referências, estruturas e significados, contextos e usos (GONZALEZ; LIMA, 2003). Dessa forma, a análise de sentimentos pode ser entendida como um problema da PLN, que tem como campo de estudo a análise de opiniões, sentimentos, avaliações, atitudes e emoções de pessoas expressas em sentenças e em textos (LIU, 2012).

Visto que o tamanho do volume de dados que atualmente se encontra disposto na internet e o aumento das possibilidades utilizando um aparelho que disponha de internet, as formas de compra e venda atuais são diferentes como as que eram utilizadas alguns anos atrás. O consumidor tem cada vez mais facilidade de adquirir algo, e está mais preocupado com a qualidade do que está adquirindo, seja este um produto eletrônico ou algum tipo de serviço. Adicionando a isso, existem ainda as empresas, que estão cada vez mais preocupados com o pós-atendimento, e como ela está sendo avaliada pelos seus clientes, fazendo com que essas marcas também

Yago A. Santos
Instituto de Biotecnologia, Universidade Federal de Catalão, Catalão, Goiás, Brasil.
e-mail: yago.alves@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.

busquem formas de estar a par do que o seu consumidor tem a dizer sobre o que acabou de adquirir.

Motivado pelo elevado número de usuários de aparelhos celulares e smartphones no país, o trabalho aqui proposto tem como objetivo realizar um desenvolvimento pautado na avaliação de técnicas para análise de sentimento e que serão descritas neste artigo.

Nas próximas seções deste artigo, será abordado os principais trabalhos relacionados com a área de análise de sentimento (seção 2), o corpus e as ferramentas utilizadas no desenvolvimento do trabalho serão listados e descritos na seção 3. Na seção 4 é descrito o desenvolvimento do trabalho, na seção 5 os resultados do experimentos e análise sobre esses resultados será mostrado, e na seção 6, uma conclusão sobre todo o trabalho será apresentado.

2 Trabalhos Relacionados

A análise de sentimentos é um problema conhecido e discutido em diversos trabalhos da área de PLN. Nesta seção, será abordado os principais trabalhos de referência da área e também trabalhos atuais que discutem sobre a mineração de opiniões e seus principais desafios.

O trabalho de Hu e Liu (2004) busca explorar o problema de geração de resumos baseados em recursos de avaliações de clientes sobre produtos vendidos online. Não sendo necessário de um corpus como base de dados de onde retirar as avaliações que serão utilizadas no treinamento e testes para realizar a tarefa, aqui o autor utiliza como entrada do sistema um nome de produto e uma página da web, onde a saída é o resumo das revisões, que por sua vez é obtido através de 3 etapas: 1) recursos do produto de mineração comentados pelos clientes; 2) identificação da sentença de opinião, utilizando como ferramenta o analisador linguístico NLProcessor[1] e o WordNet[2] ; 3) a orientação de cada sentença de opinião é identificada produzindo o resumo final.

Turney (2002) apresenta em seu trabalho uma classificação de avaliações feita com um algoritmo de aprendizado de máquina não supervisionado PMI-IR (CHURCH; HANKS, 1990). O algoritmo utiliza de revisões/avaliações escritas como entrada e produz uma classificação como saída também através de 3 etapas: 1) extrai frases contendo adjetivos ou advérbios do corpus Epinions[3]; 2) estima a orientação das frases extraídas, usando o algoritmo PMI-IR; 3) é realizado o cálculo da orientação semântica média das frases na avaliação fornecida e classifica a avaliação como recomendada se a média for positiva e caso contrário não recomendada.

Um dos trabalhos mais recentes da área para o português, o projeto de Avanço (2015) é dividido em três principais etapas, na primeira ele buscar por algoritmos para classificação de opiniões em textos da web escritos para o português brasileiro, no segundo momento o autor busca implementar classificadores de opinião e avaliar seus desempenhos na classificação de determinado tipo de opinião, e por último desenvolver métodos e ferramentas para a normalização de conteúdos gerados pelos usuários. Neste trabalho, o autor define como domínio a avaliação de produtos, utiliza de métodos para classificação de opinião a nível de documento, classificadores *baseline*, baseado em léxico, baseados em aprendizado de máquina, híbrido e classificadores já

[1] NLProcessor - Kit de ferramentas de análise de texto. http://www.infogistics.com/textanalysis.html

[2] https://wordnet.princeton.edu

[3] http://www.epinions.com

citados anteriormente combinados com conhecimento semântico obtido pela modelagem de palavras em um espaço vetorial. O corpus utilizado foi o do Buscapé[4], onde no momento de avaliar a aderência dos resultados obtidos a esse corpus de referência, foi utilizado outros três corpus, o ReLi[5] utilizado para avaliação de livros, o do Mercado Livre[6], e uma segunda versão do corpus Buscapé.

No trabalho desenvolvido por Machado (2018), o mesmo busca implementar, analisar, melhorar e criar métodos de análise de sentimentos baseada em aspectos para textos em português. Focado nos métodos não supervisionados, que neste caso se dá através de implementação e avaliação de métodos de análise de sentimentos baseada em aspectos, o autor utiliza de métodos de frequência, baseadas em substantivos, sintagmas nominais e também por modelo de semântica distribucional Word2Vec (MIKOLOV et al., 2013) para identificar os aspectos. Na fase de atribuição de sentimentos, foram utilizados dos léxicos de sentimentos OpLexicon (SOUZA; VIEIRA, 2011), Sentilex (SILVA et al., 2010), LIWC (BALAGE FILHO; PARDO; ALUÍSIO, 2013) e também o LexReLi, desenvolvido pelo próprio autor. Todo o trabalho foi desenvolvido utilizando do córpus ReLi, e utilizando também do analisador morfológico do módulo *nlpnet*[7].

3 Corpus e Recursos

Um corpus pode ser entendido de forma simplificada como um conjunto de texto que representa informações sobre determinado domínio. Sánchez (1995) define corpus como:

> 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 da 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 vários e úteis para a descrição e análise. (SÁNCHEZ, 1995, p. 8-9).

O corpus utilizado neste trabalho foi desenvolvido por Hartmann et al. (2014), sendo composto por avaliações de produtos reunidos do site Buscapé . O corpus reúne 85.910 análises de produtos, com 4.097.905 tokens (palavras) e 68.633 tipos (tipos de produtos, por exemplo, TV, celulares e smartphones, câmeras digitais, perfumes, jogos). Em geral, as categorias de produtos mais frequentes nas análises são (em ordem decrescente de frequência): TV, telefones celulares e smartphones, câmeras digitais, perfumes, jogos, condicionadores de ar, notebooks e tablets. Depois de remover palavras irrelevantes, números e pontuação, se obteve 63.917 tipos.

O domínio utilizado nesse trabalho, corresponde a Celular e Smartphone. As avaliações desses produtos estão estruturadas em pastas e divididas no esquema de estrelas (utilizado para determinar a polaridade de cada *review*) de 0 a 5, reunindo todas as avaliações que receberam 0 estrelas no site Buscapé em uma subpasta, as que receberam 1 estrela em outra subpasta, e assim para todas as avaliações, de 0 a 5 estrelas, onde o número de análises dentro de cada uma dessas pastas não é padrão, variando entre 218 arquivos que receberam 0 estrelas a 5.768 análises que receberam 5 estrelas.

[4] https://www.buscape.com.br/

[5] https://www.linguateca.pt/Repositorio/ReLi/

[6] https://www.mercadolivre.com.br/

[7] http://nilc.icmc.usp.br/nlpnet/

Dentro das técnicas implementadas, existem aquelas que se dão por meio de análise de textos processados por um etiquetador morfossintático, onde a ferramenta trata de classificar as palavras em categorias gramaticais, com base na função que desempenham na sentença e no contexto em que são utilizadas, buscando identificar padrões em que estão presentes adjetivos, substantivos, advérbios, entre outras classes gramaticais. Os etiquetadores utilizados na análise morfossintática foi o TreeTagger[8], NLTK[9], Spacy[10], e o Palavras[11]. O normalizador Enelvo[12] foi utilizado para limpar os textos e melhorar a precisão das tarefas realizadas pelos *taggers*. Sobre os léxicos mais utilizados na área e que poderiam melhor atender a nossa demanda, foi possível definir que seria utilizado no trabalho os seguintes léxicos:

- SentiLex: Conta com 82.347 entradas divididas entre expressões idiomáticas (34.700), verbos (29.504) e substantivos (1.280) classificados em negativos (53.973), positivos (20.670) e neutros (7.704). A polaridade foi atribuída no formato de graduação numérica com valores entre -10 e 10, onde zero é considerado neutro.
- OpLexicon: A ferramenta contém um total de 32.191 entradas, distribuídas entre *emoticons* (66), termos com *hashtag* (471), adjetivos (24.475), verbos (6.889) e expressões (290) classificados em negativos (14.569), neutras (9002) e positivos (8620).
- SentiWordNet-PT-BR: Esse recurso lexical em português é oriundo do escrito em inglês por Bo Pang (2008), e conta com 117.374 entradas vindas de anotações automáticas de todos os *synsets* do WordNet 3.0. Cada *synsets* tem em média três palavras e conta também com três pontuações numéricas Obj(s), Pos(s) e Neg(s), onde cada uma das três pontuações varia entre 0.0 e 1.0 e a sua soma total vale 1.0 para cada *synsets*.
- LIWC: Foi obtido via tradução do dicionário original em inglês (PENNEBAKER; FRANCIS; BOOTH, 2001) e é composto por 127.160 entradas, na qual as mesmas não são morfologicamente categorizadas, e não existe classe para os neutros. O léxico é dividido em positivos (12.878), negativos (15.115) e ainda em palavras de negação (19), saúde (7.003), dinheiro (5.353), corpo (4.766), amizade (679), família (96), entre outras, totalizando 64 categorias.
- Concatenados: Como uma segunda forma de utilizar os léxicos disponíveis para serem aplicados no trabalho, foi feito ainda de forma concatenada a junção dos léxicos citados anteriormente a fim de obter melhores resultados na tarefa de avaliação (MACHADO, 2018).

A mineração de opinião utiliza de métodos para realizar a avaliação de sentimento de um conjunto de dados, tais métodos podem ser caracterizados segundo os paradigmas de classificação supervisionada, guiada pelo uso de um léxico, supervisionada híbrida e baseada em grafos (SILVA, 2016).

No método de classificação supervisionado, onde se utiliza de um classificador de aprendizado de máquina. Cavalcante (2017) diz que a técnica “consiste na utilização de um conjunto de dados anteriormente rotulados para prever rótulos de dados futuros”. Esse se difere do método baseado em léxico pelo fato de que neste, por sua vez, “utilizam listas de palavras anotadas por polaridade ou pontuação de polaridade para determinar a pontuação geral de opinião de um determinado texto” (GIACHANOU; CRESTANI, 2016).

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

[9] http://www.nltk.org/

[10] https://spacy.io/

[11] https://portal.findresearcher.sdu.dk/en/publications/the-parsing-system-palavras-automatic-grammatical-analysis-of-por

[12] https://thalesbertaglia.com/enelvo/sobre/

O método de classificação híbrido utiliza da junção de recursos de aprendizado de máquina com as abordagens baseadas no léxico, onde Liu (2012) explica a técnica como sendo um processo que produz uma variável indicadora que é extraída de uma distribuição multinominal governada por um conjunto de parâmetros. No fim, a variável indicadora determina se uma palavra na frase é uma palavra de aspecto, uma palavra de opinião ou uma palavra de fundo.

Diferente dos métodos tratados anteriormente, no qual analisavam características léxicas, semânticas e sintáticas que podem ser encontradas em qualquer tipo de fonte textual, a abordagem baseada em grafos (Graph-Based), segundo Giachanou e Crestani (2016), é utilizada por pesquisadores sob a suposição de que as pessoas influenciam umas às outras, sendo assim, explorando propriedades especificas da própria rede social.

4 Desenvolvimento

O desenvolvimento do trabalho corresponde as fases de pré-processamento do corpus e identificação de sentimento, que é realizado na aplicação das técnicas.

4.1 Técnicas

A análise de sentimentos baseada em aspecto é direcionada a analisar textos contendo opinião e buscar identificar e relacionar sentimentos a aspectos de uma determinada entidade (MACHADO, 2018). Nesse trabalho foi utilizado dois tipos de extração entre as que se tem disponíveis atualmente para realizar a atividade, a extração de aspectos utilizando substantivos e da extração de aspectos utilizando *embedding* (COSTA; PARDO, 2020).

Na extração utilizando substantivos, a técnica se deu buscando pelos substantivos do córpus do Buscapé, dos *reviews* de celular e smartphones, onde esses substantivos foram extraídos baseados na classe que o etiquetador morfossintático TreeTagger etiquetou a palavra, adicionando o termo a um léxico caso ele tenha sido classificado como substantivo e salvando o conjunto de palavras em um arquivo.

A segunda forma abordada no trabalho para realizar a extração de aspectos foi utilizando *embedding*. Conhecida por utilizar representação vetorial de palavras, a técnica busca representar cada palavra com um ponto, ou um vetor, em um espaço semântico multidimensional (COSTA; PARDO, 2020). Resumidamente, palavras com sentido parecido se alocam no espaço próximo uma da outra, por exemplo, a palavra “bom” ao aplicar a um método de *embedding*, a mesma irá se encontrar dentro de um vetor com posição próxima de outras palavras com o sentido/significado parecido, como “fantástico”, “ótimo”, “maravilhoso”.

Dentre os métodos de *embedding* existe o Word2Vec . Tal método faz parte da família de métodos utilizados na representação de palavras como vetores, gerando vetores densos de onde será utilizado de palavras “sementes” para extrair outros aspectos que podem ser encontrados dentro do domínio que está sendo utilizado na análise de sentimento, ou seja, palavras são utilizadas como entrada no método e a saída serão palavras de sentido parecidos com a palavra de entrada.

A abordagem baseada em léxico, assim como outros tipos de abordagens, tem diversas maneiras de realizar a implementação das técnicas. Nesta é considerado um léxico de sentimento para realizar a avaliação dos *reviews* e identificar o sentimento de uma sentença. As técnicas aplicadas neste trabalho foram basicamente três:

Posição do adjetivo (Técnica 1): Parecida com a forma de avaliação proposta por Freitas e Vieira (FREITAS; VIEIRA, 2015) e pela técnica identificada como “Somente Adjetivos” por Machado (2018), a técnica consiste em analisar se existe um adjetivo antes ou depois do substantivo/aspecto, nessa ordem, após analisado é atribuído a polaridade do adjetivo encontrado ao substantivo ou ao aspecto.

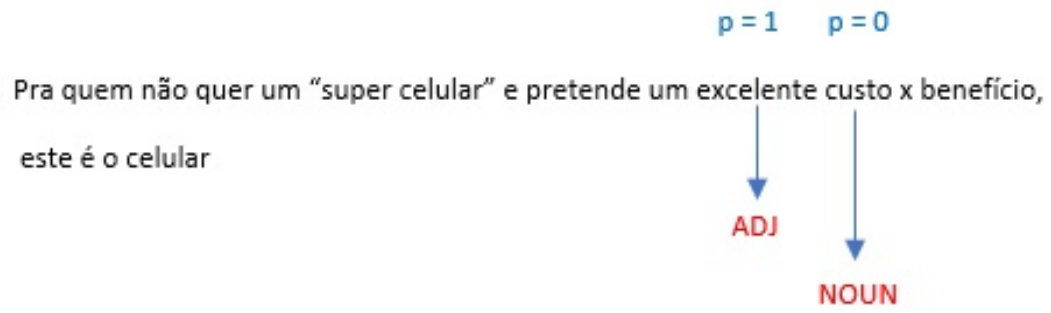

Figura 1: frase com classe gramatical atribuída.

Na Figura 1 vemos uma frase na qual “custo” é uma palavra classificada como substantivo (NOUN), a palavra “excelente” está classificada como adjetivo (ADJ). Suas polaridades iniciais são respetivamente 0 e 1, dessa forma, como já citado nessa seção, quando a técnica identifica um adjetivo antes ou depois de um substantivo/aspecto, a polaridade do adjetivo é atribuída ao substantivo, sendo assim, a polaridade da palavra “custo” passaria a ser 1. Dando continuidade a técnica, é feita a mesma análise para todas as palavras das sentenças, que seja um aspecto/substantivo e que venha acompanhado por um adjetivo, ao final, é realizado a somatória do valor da polaridade de cada palavra de sentimento do *review* para poder definir se a polaridade do mesmo é positiva ou negativa.

Posição do adjetivo + polaridade das palavras (Técnica 2): Nomeada por Machado (2018) como “Preferência aos adjetivos”, a técnica se trata de uma combinação da anterior, onde é realizado uma busca por adjetivos antes e/ou depois do substantivo/aspecto, e caso o algoritmo não encontre adjetivos com polaridade, o mesmo analisa as polaridades de todas as palavras da sentença. Diferente da técnica anterior, após realizar a análise de todas as palavras e verificar que não tem nenhum substantivo seguido de adjetivo, a técnica entra em uma segunda possibilidade, que é a de apenas somar todas as polaridades das palavras de sentimentos encontradas no texto para realizar a avaliação do *review*.

CBL (Técnica 3): Diferente das técnicas anteriores que utilizava de etiquetadores morfossintáticos para classificar as palavras e posteriormente em conjunto com isso realizar a avaliação do texto, o classificador baseado em léxico desenvolvido por Avanço (2015), segue a proposta de Taboada et al. (2011) partindo do princípio de que palavras de sentimento possuem uma polaridade, e que caso essas palavras ocorram em um contexto de negação, intensificação, ou redução, as respectivas polaridades são modificadas. A técnica se dá buscando palavras de sen-

timento (palavras com polaridade atribuída) que após encontrada é verificado se a mesma está acompanhada de alguma das palavras definidas anteriormente como de negação, intensificação ou redução. No final da técnica é realizado o cálculo da orientação semântica geral por meio das somas das polaridades encontradas. Nesta técnica, o algoritmo basicamente busca por uma palavra de sentimento, quando encontrada, as 3 palavras que aparecem após a palavra de sentimento são extraídas da sentença e verificada se dentre essas 3 palavras aparecem algum termo de negação, intensificação ou redução. Conforme a combinação que aparece entre os termos (negação seguido de intensificação, negação seguido de redução, ou outra combinação possível), o algoritmo realiza o cálculo da polaridade do *review*.

Todas as técnicas baseadas em um léxico que acabaram de ser mencionadas foram desenvolvidas utilizando busca por substantivos (aspectos) no momento que a mesma estava fazendo a avaliação do texto, e implementadas de uma segunda forma utilizando dos aspectos extraídos em um primeiro momento e reservados para uso posterior, e que já foram citados nesta seção como se deu essa extração.

Uma segunda característica que todas as técnicas implementadas nesse trabalho carregam, é que todas utilizam da regra linguística da negação a fim de buscar por partículas que possam alterar a polaridade do termo que está sendo trabalhado no algoritmo. Entre as possibilidades de se tratar negação, existe a forma onde a partícula negativa aparece antes do aspecto (i), existe o caso onde a partícula negativa aparece depois do aspecto (ii), e a dupla negação (iii), onde a palavra conta com uma partícula negativa antes e uma depois.

5 Experimentos e Análises

Tendo como a abordagem escolhida para dar andamento nas aplicações do trabalho o paradigma de classificação baseada em um léxico, várias técnicas foram implementadas e avaliadas com o objetivo de obter o melhor resultado na atividade de analisar corretamente o sentimento atribuído a uma determinada avaliação. Do corpus que está sendo utilizado, foram extraídos 2.242 *reviews* positivos e a mesma quantidade de negativos, onde desse número 75% do corpus foi empregado no treinamento das técnicas, totalizando 1.682 avaliações positivas e a mesma quantidade de avaliações negativas, onde os outros 25% dessas análises foram reservados para testes, totalizando 560 *reviews*.

A forma de avaliação utilizada no trabalho se deu comparando os sentimentos identificados após as técnicas serem aplicadas nos *reviews* do corpus Buscapé, com a polaridade extraída do mesmo. Utilizando da forma de avaliação semelhante a utilizada nos workshops SemEval (PONTIKI et al., 2015), foram empregadas das métricas de cobertura, precisão, medida-f e acurácia para realizar a avaliação dos experimentos, no qual foram realizados testes para cada uma das técnicas utilizando mesmo corpus, tanto para treinamento, quanto para testes das técnicas mencionadas neste trabalho.

Para que fosse possível realizar os cálculos, classes foram definidas e no momento da avaliação e associadas a cada tipo de polaridade encontrada após aplicar cada tipo de técnica. Nomeado como TP (*true positive*) os reviews identificados corretamente como positivos, FP (*false positive*) os classificados como positivos sendo negativos, TN (*true negative*) reviews negativos que foram classificados corretamente como tal, e FN (*false negative*) sendo os que foram classi-

ficados como negativo sendo positivos. Dessa forma, as fórmulas utilizadas como métrica foram aplicadas após cada técnica de avaliação ter sido executada.

A primeira técnica abordada será a que utiliza de avaliação baseada em léxico com aspectos extraídos anteriormente por substantivos e *embedding*, utilizando do etiquetador Treetagger, no qual foi o único escolhido para aplicar nessa técnica no momento que o trabalho se encontrava por se tratar de um dos etiquetadores que utiliza o conjunto de *tags* com base nas recomendações do Grupo Consultivo de Especialistas em Padrão de Engenharia de Idiomas (EAGLES) . Nesse caso, a técnica posição de adjetivo (I) e posição de adjetivo + polaridade das palavras (II) foram as únicas que puderam ser aplicadas utilizando de extração de aspecto, entre as abordadas nesse trabalho, por se tratar de técnicas que realizam a avaliação buscando por um termo com tal características.

Na segunda leva de testes, foram aplicadas as técnicas utilizando de todos os léxico de sentimento abordados, variando entre os etiquetadores disponíveis que foram citados na seção anterior, acrescido de que nesse segundo momento, não foi utilizado de aspectos extraídos anteriormente, visto que toda identificação de termo necessários para a realização das técnicas foi realizado no momento da avaliação pelo etiquetador que estava sendo empregado.

Na terceira técnica de avaliação baseada em léxico, a mesma não utiliza de busca por qualquer tipo de classe atribuída por um etiquetador morfossintático, em vez disso ela busca por palavras com polaridade atribuída, definindo uma janela de tamanho 3[13] logo após a palavra com polaridade, e utilizando disso para realizar a técnica.

Após aplicar e analisar as diferentes abordagens que foram citadas no trabalho, foi possível identificar que certas ferramentas e técnicas se destacaram para com o domínio e o objetivo da pesquisa. Nas técnicas que utilizaram de extração de aspectos, a que se saiu melhor foi a Posição de Adjetivo em conjunto com Polaridade das Palavras utilizando de aspectos extraídos por Substantivos, unido do etiquetador TreeTagger com o léxico Sentilex. Já nas técnicas que não utilizou do recurso de extração de aspectos, na técnica 1 o melhor desempenho foi obtido unindo do léxico SentiLex com o etiquetador NLTK, na técnica 2 utilizando do NLTK em conjunto com o SentiLex, e na técnica 3 obteve o melhor resultado *reviews* classificados com o SentiWordNet (Tabela 1).

Técnica	Etiquetador	Léxico	Acurácia	Medida-F
Técnica 1	NLTK	SentiLex	77.75%	0.754
Técnica 2	NLTK	SentiLex	75.06%	0.741
Técnica 2 + Aspectos extraídos por Substantivos	TreeTagger	SentiLex	72.81%	0.764
Técnica 3	-------	SentiWordNet-PT-BR	70.05%	0.610

Tabela 1: Melhores resultados por cada tipo de técnicas aplicadas.

[13] Valor definido através de testes empíricos

6 Conclusão

Nesse trabalho foram abordadas das principais técnicas de análise de sentimentos com o intuito de observar qual das abordagem melhor se encaixa na atividade de realizar a identificação correta dos sentimentos associados a uma avaliação do domínio de produtos eletrônicos, mais especificamente celular e smartphone, a fim de construir uma plataforma capaz de receber como entrada o nome do produto e como saída o usuário ter uma avaliação do produto em si e das suas principais características que foram anteriormente avaliado por outros usuários.

Como proposta de trabalhos futuros pretende-se aplicar demais técnicas de mineração de opinião já conhecidas e que não foram abordadas no trabalho, como aprendizado de máquina, com o objetivo de comparar qual o melhor método de avaliação para o domínio proposto, além da criação da plataforma em si, que fará a uso da abordagem que tiver melhor desempenho nos testes realizados com o conjunto de corpus e métricas de avaliação definidas.

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.

BO PANG, Lillian Lee. Opinion mining and sentiment analysis. In: FOUNDATIONS and Trends® in Information Retrieval. [S.l.: s.n.], 2008.

CAVALCANTE, Paulo Emílio Costa. Um dataset para análise de sentimentos na língua portuguesa. Universidade Federal da Paraíba, 2017.

CHURCH, Kenneth; HANKS, Patrick. Word association norms, mutual information, and lexicography. **Computational linguistics**, v. 16, n. 1, p. 22–29, 1990.

COSTA, Raul Wagner Martins; PARDO, Thiago Alexandre Salgueiro. Métodos baseados em léxico para extração de aspectos de opiniões em português. In: SBC. ANAIS do IX Brazilian Workshop on Social Network Analysis and Mining. [S.l.: s.n.], 2020. p. 61–72.

FREITAS, Larissa A de; VIEIRA, Renata. Exploring resources for sentiment analysis in Portuguese language. In: IEEE. 2015 Brazilian conference on intelligent systems (BRACIS). [S.l.: s.n.], 2015. p. 152–156.

GIACHANOU, Anastasia; CRESTANI, Fabio. Like it or not: A survey of twitter sentiment analysis methods. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 49, n. 2, p. 1–41, 2016.

GONZALEZ, Marco; LIMA, Vera Lúcia Strube. Recuperação de informação e processamento da linguagem natural. In: XXIII Congresso da Sociedade Brasileira de Computaçao. [S.l.: s.n.], 2003. v. 3, p. 347–395.

HARTMANN, Nathan et al. A Large Corpus of Product Reviews in Portuguese: Tackling Out-Of-Vocabulary Words. In: LREC. [S.l.: s.n.], 2014. p. 3865–3871.

HU, Minqing; LIU, Bing. Mining and summarizing customer reviews. In: PROCEEDINGS of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. [S.l.: s.n.], 2004. p. 168–177.

LIU, Bing. Sentiment analysis and opinion mining. **Synthesis lectures on human language technologies**, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–167, 2012.

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.

MIKOLOV, Tomas et al. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

PENNEBAKER, James W; FRANCIS, Martha E; BOOTH, Roger J. Linguistic inquiry and word count: LIWC 2001. **Mahway: Lawrence Erlbaum Associates**, v. 71, n. 2001, p. 2001, 2001.

PONTIKI, Maria et al. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In: PROCEEDINGS of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Denver, Colorado: Association for Computational Linguistics, jun. 2015. p. 486–495. DOI: 10.18653/v1/S15-2082. Disponível em: ¡https://www.aclweb.org/anthology/S15-2082¿.

SÁNCHEZ, Aquilino. Definición e historia de los corpus. **Cumbre: Corpus lingüístico del Español contemporáneo. Madrid: SGEL**, p. 7–24, 1995.

SILVA, Mário J et al. Automatic expansion of a social judgment lexicon for sentiment analysis, 2010.
SILVA, Nadia Felix Felipe da. **Análise de sentimentos em textos curtos provenientes de redes sociais**. 2016. Tese (Doutorado) – Universidade de São Paulo.
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.
TABOADA, Maite et al. Lexicon-based methods for sentiment analysis. **Computational linguistics**, MIT Press, v. 37, n. 2, p. 267–307, 2011.
TURNEY, Peter D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. **arXiv preprint cs/0212032**, 2002.