ANÁLISE DE SENTIMENTO EM NÍVEL DE ASPECTO COM BASE EM ONTOLOGIAS DE DOMÍNIO E ANÁLISE DE DEPENDÊNCIA SINTÁTICA

FRANCISCO DIAS FRANCO[1]; LARISSA ASTROGILDO DE FREITAS[2]; ULISSES BRISOLARA CORRÊA[3]

[1] *Universidade Federal de Pelotas – fdfranco@inf.ufpel.edu.br*
[2] *Universidade Federal de Pelotas – larissa.freitas@ufpel.edu.br*
[3] *Universidade Federal de Pelotas – ulisses@inf.ufpel.edu.br*

1. INTRODUÇÃO

A área responsável por analisar avaliações e extrair o sentimento expresso por ela é chamada de Análise de Sentimento (do inglês, *Sentiment Analysis* - SA). A SA é responsável por classificar o sentimento que está atrelado a uma opinião (como, uma avaliação de um produto), tendo em vista emoção (por exemplo: felicidade, tristeza, medo e raiva), polaridade (por exemplo: positivo e negativo) ou intensidade (por exemplo: muito positivo e muito negativo) de um aspecto (por exemplo: preço e durabilidade de um produto) (LIU, 2020).

No entanto, em idiomas de poucos recursos, como o português, existem poucos modelos e há uma quantidade escassa de *datasets* abertos com os quais modelos podem ser treinados e testados (CIERI et al., 2016). Da mesma forma que precisamos de uma variedade de recursos linguísticos para entender, compreender e expressar sentimentos e opiniões, os modelos também precisam desses recursos para sintetizar informações (CHAPELLE, 2020).

O intuito desse trabalho é realizar a expansão da pesquisa desenvolvida por Freitas (2015). O diferencial do presente trabalho é o uso da análise de dependência sintática na identificação de termos opinativos relativos aos aspectos encontrados nas *reviews* de hotéis e o uso de diferentes léxicos de sentimentos. Para isso, iremos utilizar o analisador de dependência sintática da biblioteca spaCy (HONNIBAL et al., 2020).

2. METODOLOGIA

A metodologia utilizada será uma adaptação da metodologia desenvolvida por Freitas (2015). Ela possui as seguintes etapas:

1. **Pré-processamento:** Essa etapa consiste em aplicar as etapas de *tokenization*, *lemmatization* e *POS tagging* no *dataset*.
2. **Identificação das características:** Essa etapa consiste em identificar os aspectos (conceitos ou características) nas opiniões pré-processadas, isso é feito com o uso do HOntology (CHAVES; FREITAS; VIEIRA, 2012).
3. **Identificação da Polaridade:** Essa etapa consiste em identificar a polaridade das opiniões pré-processadas que contêm características, isso é feito com o uso da árvore de dependência sintática e de léxicos de sentimento.
4. **Sumarização dos Resultados:** Essa etapa consiste em apresentar as polaridade referentes aos aspectos encontrados na opinião analisada.

Ainda, foi necessário realizar o tratamento dos léxicos de sentimento. O *dataset* utilizado e as etapas de tratamento dos léxicos de sentimento são descritas nas próximas seções.

2.1 *DATASET*

O *dataset*[1] que será utilizado durante presente trabalho foi composto por *reviews* de usuários do TripAdvisor[2], sendo que cada *review* pode, ou não, conter uma ou mais opiniões. Ele é composto por 3.111 amostras, provenientes de 847 avaliações sobre o setor hoteleiro. Sendo 2.112 amostras positivas, 527 amostras negativas e 472 amostras neutras. Das 3.111 amostras, 255 contém aspectos (compostos por mais de uma palavra) e 2856 contém aspectos simples (compostos por uma única palavra).

2.2 TRATAMENTO DOS LÉXICOS DE SENTIMENTO

Ao total foram testados doze Léxicos de Sentimento: AffectPT-br[3] (com e sem Wildcard Expansion), EmoLex[4], LeIA[5], LIWC[6], OpLexicon (SOUZA et al., 2011; SOUZA; VIEIRA, 2012), OntoPT[7], Reli-Lex[8], SentiLex[9], SentiWordNet-PT-BR[10], UNILEX[11] e WordNetAffectBR[12].

Em tese, o tratamento dos léxicos[13] consiste em uma padronização de cada léxico, de forma a facilitar o acesso à informação e a implementação do sistema proposto. Foi gerado um dicionário referente à cada léxico utilizado.

3. RESULTADOS E DISCUSSÃO

A Tabela 1 mostra o resultado do modelo para cada um dos léxicos de sentimento, utilizando as métricas: acurácia (*Accuracy*), precisão (*Precision*), revocação (*Recall*) e medida-f (*F1-score*). Sendo que todos os resultados, exceto acurária, possuem as médias *macro*, *micro* e *weighted* (poderada).

---

[1]Mais informações sobre o *dataset* utilizado podem ser obtidos através do link <https://encurtador.com.br/mnqW8>.

[2]O TripAdvisor é o maior site de viagens do mundo, possuindo mais de 859 milhões de avaliações e opiniões sobre 8,6 milhões de acomodações, restaurantes, experiências, companhias aéreas e cruzeiros. Dessa forma, ele se torna uma importante fonte de informações para treinamento de modelos de SA para serviços de hospedagens.

[3]Disponível em: <https://github.com/LaCAfe/AffectPT-br>.

[4]Disponível em: <https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>.

[5]Disponível em: <https://github.com/rafjaa/LeIA>.

[6]Disponível em: <https://sites.icmc.usp.br/sandra/LIWC/LIWC2007_Portugues_win.dic>.

[7]Disponível em: <http://ontopt.dei.uc.pt/index.php?sec=download_ontopt>.

[8]Disponível em: <https://www.linguateca.pt/Repositorio/ReLi/>.

[9]Disponível em: <https://b2share.eudat.eu/records/93ab120efdaa4662baec6adee8e7585f>.

[10]Disponível em: <https://github.com/Pedro-Thales/SentiWordNet-PT-BR>.

[11]Disponível em: <https://dicionariounilex.wixsite.com/unilex>.

[12]Disponível em: <https://www.inf.pucrs.br/linatural/wordpress/recursos-e-ferramentas/wordnetaffectbr/>.

[13]Mais informações sobre os recursos léxicos e tratamento dos léxicos podem ser obtidas através do link: <https://encurtador.com.br/bGJ48>.

Tabela 1: Resultados do modelo para cada um dos léxicos de sentimento.

Metric / Lexicon	lengths	Accuracy	Precision (macro)	Precision (micro)	Precision (weighted)	Recall (macro)	Recall (micro)	Recall (weighted)	F1-score (macro)	F1-score (micro)	F1-score (weighted)
AffectPT-br	1135	0,53	**0,63**	0,53	**0,64**	0,41	0,53	0,53	0,37	0,53	0,47
AffectPT-br c/WE	17617	0,56	**0,62**	0,56	**0,64**	0,44	0,56	0,56	0,42	0,56	0,52
EmoLex	11231	0,49	0,55	0,49	**0,62**	0,40	0,49	0,49	0,34	0,49	0,39
LeIA	5779	0,56	**0,60**	0,56	**0,64**	0,44	0,56	0,56	0,41	0,56	0,51
LIWC	27491	0,51	0,53	0,51	0,57	0,41	0,51	0,51	0,38	0,51	0,46
OpLexicon	32119	**0,57**	0,56	**0,57**	0,60	**0,47**	**0,57**	**0,57**	**0,47**	**0,57**	**0,55**
OntoPT	14039	0,54	0,52	0,54	**0,62**	0,46	0,54	0,54	0,44	0,54	0,51
Reli-Lex	345	0,54	**0,68**	0,54	**0,67**	0,43	0,54	0,54	0,39	0,54	0,48
SentiLex	7010	0,56	**0,63**	0,56	**0,66**	0,45	0,56	0,56	0,43	0,56	0,51
SentiWordNet	6007	0,53	0,50	0,53	**0,61**	0,45	0,53	0,53	0,42	0,53	0,49
UNILEX	3845	0,54	0,53	0,54	**0,61**	0,44	0,54	0,54	0,41	0,54	0,50
WordNetAffectBR	289	0,44	**0,72**	0,44	**0,69**	0,34	0,44	0,44	0,22	0,44	0,27

Como podemos notar na Tabela 1, o OpLexicon possui quase as métricas com mais relevante, embora as métricas *Precision* (*macro*) e/ou *Precision* (*weighted*) tenham valores maiores em todos os outros léxicos, exceto LIWC.

Isso levantou uma questão importante: a quantidade de termos (*lengths*) de um recurso léxico influencia os resultados do modelo? Se considerarmos a razão da quantidade de termos de um léxico de sentimento pela métrica avaliada, obteremos uma relação de proporcionalidade. Assim, quanto menor a quantidade de termos de um léxico de sentimento e maior é o valor da métrica avaliada, melhor é aquele recurso em relação à métrica avaliada.

Observando, por exemplo, os léxicos de sentimento OpLexicon e Reli-Lex podemos notar a discrepância entre os números de termos. Enquanto OpLexicon possui 32.119 termos, Reli-Lex possui apenas 345, isso é, OpLexicon é, aproximadamente, 93 vezes maior que Reli-Lex.

Ainda, a razão entre a acurácia e o número de termos, temos para OpLexicon $1,78.10^{-5}$, enquanto que para o Reli-Lex temos $1,57.10^{-3}$. Assim, quanto menor é a razão pior é o aproveitamento daquele léxico de sentimento para domínio analisado. Embora o domínio do Reli-Lex seja para livros, já que ele foi treinado nesse contexto, ele se mostrou bastante relevante, mesmo possuindo tão poucos termos.

4. CONCLUSÃO

A avaliação de produtos e serviços desempenha um grande impacto no desenvolvimento de negócios, pois, influencia a opinião de clientes, fazendo com que eles escolham produtos com avaliações melhores, e permite aos vendedores, e aos prestadores de serviços, conhecer as qualidades/defeitos dos seus produtos e serviços, contribuindo assim para o melhoramento na qualidade dos produtos e serviços.

A SA de avaliações de diferentes sites consome bastante tempo e a verificação da polaridade das opiniões exige um esforço maior ainda. O modelo é to-

talmente dependente de *datasets* disponíveis, quanto maior é o *dataset* disponível mais preciso e fiável é o modelo.

Para idiomas de poucos recursos temos problema com a pouca disponibilidade de *datasets* abertos e com isso os resultados nas métricas de avaliação não são tão bons quanto em comparação com idiomas com muitos recursos, como o inglês.

Dessa forma, torna-se altamente relevante trabalhos no campo de idiomas com poucos recursos, pois, precisamos desenvolver modelos mais eficientes. Isso está intimamente relacionado com a complexidade do idioma analisado. Por possuir uma complexidade elevada na língua, torna-se mais custoso desenvolver recursos linguísticos que contemplem toda a sua complexidade.

5. REFERÊNCIAS

CHAPELLE, C. A. (Ed.). **The concise encyclopedia of applied linguistics**. Hoboken, NJ: Wiley Blackwell, 2020. ISBN 9781119147367.

CHAVES, M. S.; FREITAS, L.; VIEIRA, R. Hontology: A multilingual ontology for the accommodation sector in the tourism industry. In: INSTICC. **Proceedings of the 4th International Conference on Knowledge Engineering and Ontology Development (KEOD 2012)**. Barcelona, Espanha: SciTePress, 2012. p. 149–154. ISBN 978-989-8565-30-3. ISSN 2184-3228.

CIERI, C.; MAXWELL, M.; STRASSEL, S.; TRACEY, J. Selection criteria for low resource language programs. In: **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**. [S.l.: s.n.], 2016. p. 4543–4549.

FREITAS, L. A. d. **Feature-level sentiment analysis applied to brazilian portuguese reviews**. 94 p. Tese (Doutorado em Ciência da Computação) — Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2015.

HONNIBAL, M.; MONTANI, I.; LANDEGHEM, S. V.; BOYD, A. spacy: Industrial-strength natural language processing in python. 2020.

LIU, B. **Sentiment analysis: Mining opinions, sentiments, and emotions**. 2. ed. Cambridge, United Kingdom: Cambridge university press, 2020. 1–448 p. (Studies in Natural Language Processing).

SOUZA, M.; VIEIRA, R. Sentiment analysis on twitter data for portuguese language. In: **Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language**. Berlin, Heidelberg: Springer-Verlag, 2012. (PROPOR'12), p. 241–247. ISBN 978-3-642-28884-5. Disponível em: <http://dx.doi.org/10.1007/978-3-642-28885-2_28>.

SOUZA, M.; VIEIRA, R.; BUSETTI, D.; CHISHMAN, R.; ALVES, I. M.; UNISINOS, F. D. L. Construction of a portuguese opinion lexicon from multiple resources. In: **In 8th Brazilian Symposium in Information and Human Language Technology - STIL, Mato Grosso**. [S.l.: s.n.], 2011.