EXPANSÃO DE UM FRAMEWORK DE ANÁLISE DE SENTIMENTOS EM PORTUGUÊS UTILIZANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA

LAERTE DOS SANTOS CARDOZO[1]; LARISSA ASTROGILDO DE FREITAS[2]

[1]*Universidade Federal de Pelotas (UFPel) – ldscardozo@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas (UFPel) – larissa@inf.ufpel.edu.br*

1. INTRODUÇÃO

A linguagem é a forma utilizada pelos seres humanos para se comunicarem uns com os outros. (SAVADOVSKY, 1988) define que a linguagem natural é uma comunicação complexa e inteligente entre as pessoas. Sendo assim, um dos principais modos de manifestação externa da atividade mental. Através dela, é possível expressar opiniões, sentimentos, ideias e experiências.

Com o crescente avanço na tecnologia, estão sendo criados serviços que geram grandes quantidades de dados por meio de seus usuários. Essa quantidade de informação pode conter sentimentos ou opiniões, tanto positivas quanto negativas. Conforme (PAK; PAROUBEK, 2010), quase todos os dias os usuários disponibilizam textos que expressam sentimentos ou opiniões sobre inúmeros conteúdos.

O Processamento da Língua Natural (PLN) é um campo de estudo da Inteligência Artificial (IA) que soluciona diversos problemas relacionados à geração e compreensão da linguagem natural. De acordo com (RUSSEL; NORVIG, 2013), PLN consiste no desenvolvimento de modelos computacionais para a solução de tarefas na qual dependam de dados expressos em linguagem natural.

Sendo assim, PLN possui uma aplicação que realiza a análise de textos opinativos denominada de Análise de Sentimentos (AS). Com o objetivo de examinar o texto e identificar a opinião do usuário contida no texto. Segundo (LIU, 2012), a AS analisa opiniões, sentimentos e emoções dos usuários em relação a produtos, organizações, serviços e outros.

Neste trabalho foi expandido um *framework* de AS (PALERMO, 2019) para o idioma português utilizando técnicas de Aprendizado de Máquina (AM). Essas técnicas têm alcançado excelentes resultados, conforme apresentado nos trabalhos de (JUNQUEIRA; FERNANDES, 2018) e de (KUMAR; SUBBA, 2020). O principal foco da expansão foi nos componentes de pré-processamento e nas técnicas de análise. Esse *framework* foi inicialmente desenvolvido com a abordagem léxica e a abordagem híbrida.

O *framework* escolhido para a execução desse trabalho foi escolhido por não utilizar tradutores automáticos para o processo de pré-processamento nos textos. Além disso, seus recursos foram desenvolvidos para o idioma português. Desse modo, o objetivo deste trabalho foi expandir um *framework* de AS em português utilizando técnicas de AM para a classificação dos sentimentos em textos.

2. METODOLOGIA

Para a expansão do *framework* foram acrescentados novos módulos nos componentes de pré-processamento e de técnicas de análise. A Figura 1 mostra em vermelho os módulos adicionados e em azul o módulo que recebeu novos métodos.

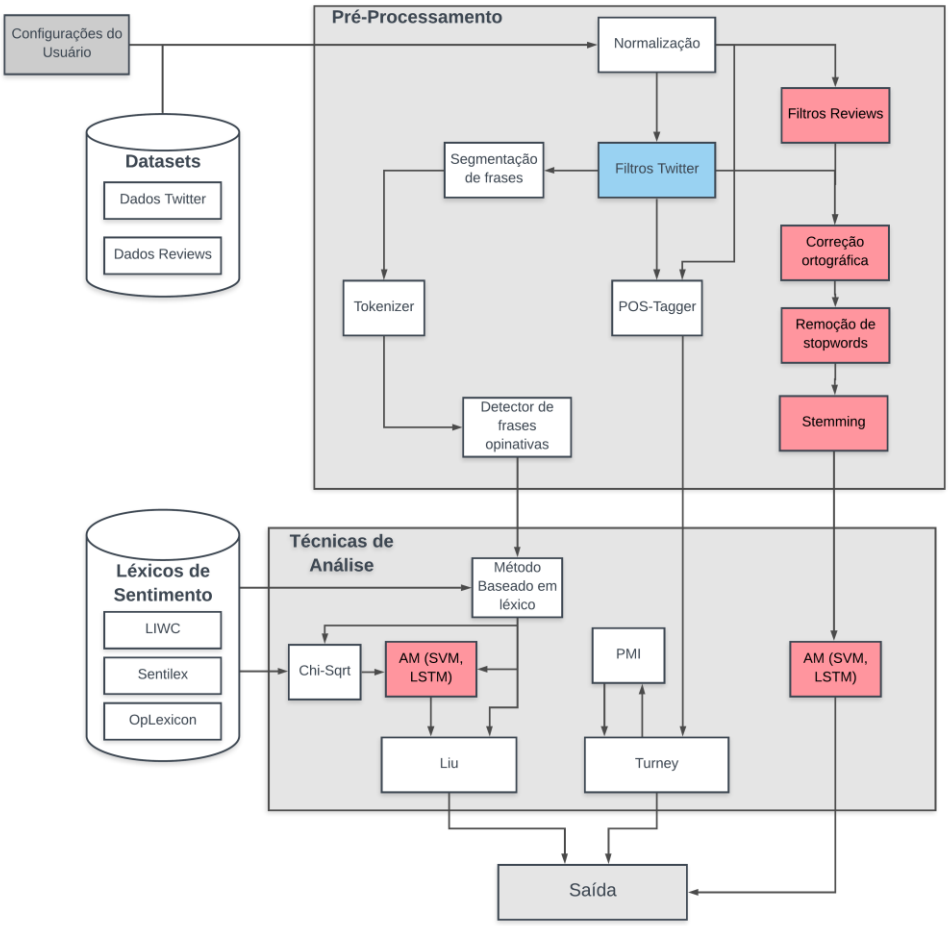

Figura 1 - Esquema da Expansão no *Framework*

O módulo de Filtros para *Reviews* é utilizado para melhorar a tarefa de pré-processamento dos textos. A correção ortográfica é utilizada para corrigir os erros ortográficos nas palavras, por exemplo, problemas de digitação. O módulo de remoção de *stopwords* remove palavras que são consideradas irrelevantes para a classificação dos sentimentos. Em princípio, são palavras que não expressam opinião, por exemplo, preposições. O *stemming* reduz as palavras para o radical, diminuindo a quantidade de palavras com mesma derivação.

As técnicas escolhidas para a tarefa de identificação dos sentimentos nos textos foram o SVM (CORTES; VAPNIK, 1995) e o LSTM (HOCHREITER; SCHMIDHUBER, 1997). A primeira técnica obteve o melhor resultado no trabalho de (PANG; LEE; VAITHYANATHAN, 2002) e é considerado o precursor em utilizar AM para AS. E a segunda técnica foi implementada no trabalho de (BRITTO; PACÍFICO, 2019), sendo a que obteve o melhor resultado entre os trabalhos mais recentes na literatura desenvolvidos para português.

Além disso, alguns métodos foram adicionados no módulo de Filtros para *Twitter* para melhorar o pré-processamento desses textos, por exemplo, remoção de *hashtags*.

Para os testes foram utilizados os corpora (conjuntos de dados/textos) para serem utilizados na análise e na classificação dos sentimentos em língua portuguesa. O primeiro corpus foi o *TweetSentBR* (BRUM; NUNES, 2018) na qual contém 15.000 *tweets* sobre programas brasileiros de televisão com três polaridades de sentimento (positivo, neutro e negativo). E o segundo corpus foi sobre produtos eletrônicos do *site Buscapé* na qual contém 10.050 *reviews* com duas polaridades de sentimento (positivo e negativo). Os dois corpora foram

balanceados para evitar que os algoritmos tenham a tendência de predizer para a polaridade com mais textos.

O processo para identificação dos sentimentos começa pelos filtros na qual remove erros/ruídos contidos nos textos. Após, corrigimos erros ortográficos nas palavras para facilitar a análise, removemos as que não expressam sentimento e as reduzimos para o radical. Desse modo, o texto está pronto para os algoritmos de AM predizer qual sentimento está expresso no texto. Com esses valores preditos, foi possível analisar a performance de cada algoritmo com determinadas métricas.

3. RESULTADOS E DISCUSSÃO

Para a obtenção da performance dos experimentos foram utilizadas as métricas de acurácia, precisão, revocação e medida F. A acurácia é definida por quanto de acerto os algoritmos tiveram com seus valores preditos. A precisão calcula o quanto de positivos foram identificados corretamente. A revocação nos diz o quanto de positivos que foram identificados corretamente entre todos os positivos. E a medida F é a média harmônica entre a precisão e a revocação.

As Tabelas 1 e 2 mostram os melhores resultados obtidos na primeira versão do *framework* e os melhores resultados parciais da expansão em cada corpus. Foram escolhidos com base na métrica de medida F, já que os resultados da primeira versão foram obtidos com os corpora desbalanceados.

Tabela 1 - Comparativo dos resultados no corpus *Buscapé*

Configuração	Acurácia	Precisão	Revocação	Medida F
Freeling (PALERMO, 2019)	83%	65%	61%	63%
SVM	80%	80%	80%	80%
LSTM	83%	83%	83%	**83%**

Tabela 2 - Comparativo dos resultados no corpus *TweetSentBR*

Configuração	Acurácia	Precisão	Revocação	Medida F
Linguakit (PALERMO, 2019)	77%	58%	60%	59%
SVM	78%	78%	78%	78%
LSTM	79%	79%	79%	**79%**

Com os resultados apresentados nas tabelas, observou-se que com base na Medida F os algoritmos de AM se mostraram bastante eficientes se comparados com as outras técnicas da primeira versão do *framework*. Além disso, verificou-se que os resultados do corpus *Buscapé* foram superiores aos resultados do corpus *TweetSentBR*. Geralmente, os textos do corpus *Buscapé* são melhores escritos do que os textos do corpus *TweetSentBR*.

4. CONCLUSÕES

Neste trabalho foi apresentado o acréscimo de uma nova abordagem utilizando técnicas de AM em um *framework* desenvolvido no idioma português. Com o objetivo de identificar os sentimentos contidos em diferentes tipos de textos (*reviews* e *tweets*) utilizando diferentes algoritmos de AM (SVM e LSTM).

Estas técnicas obtiveram melhores resultados do que a primeira versão do *framework*.

5. REFERÊNCIAS BIBLIOGRÁFICAS

BRITTO, L.F.S.; PACÍFICO, L.D.S. Sentiment Analysis for Mobile App Reviews in Brazilian Portuguese. In: **ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTAÇÃO (ENIAC 2019)**. Salvador, 2019. **Anais...** SBC, 2019. p.1-12.

BRUM, H.B.; NUNES, M.G.V. Building a sentimento corpus of tweets in brazilian portuguese. In: **ELEVENTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2018)**. Miyazaki, 2018. **Anais...** European Language Resources Association(ELRA), 2018.

CORTES, C.; VAPNIK, V. Support-Vector Networks. **Machine Learning**, USA, v.20, n.3, p.273-297, 1995.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. **Neural Computation**, USA, v.9, n.8, p.1735-1780, 1997.

JUNQUEIRA, K.T.; FERNANDES, A. Análise de Sentimento em Redes Sociais no Idioma Português com Base em Mensagens do Twitter. In: **COMPUTER ON THE BEACH**, 2018. **Anais...**, 2018. p.681-690.

KUMAR, V.; SUBBA, B. A TfidfVectorizer and SVM based sentimento analysis framework for text data corpus. In: **NATIONAL CONFERENCE ON COMMUNICATIONS (NCC)**. 2020. **Anais...**, 2020. p.1-6.

LIU, B. **Sentiment analysis and opinion mining**. USA: Morgan & Claypool Publishers, 2012.

PAK, A. PAROUBEK, P. Twitter as a corpus for sentiment analysis and opinion mining. In: **SEVENTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC'10)**. Valletta, 2010. **Anais...** European Language Resources Association (ELRA), 2010. p.1320-1326.

PALERMO, F.T.T. **Framework para Análise de Sentimentos em Nível de Documento para o Português**. 2019. Monografia (Graduação em Engenharia de Computação) - Curso de Bacharelado em Engenharia de Computação, Universidade Federal de Pelotas.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up? Sentiment Classification using Machine Learning Techniques. In: **CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP 2002)**, 2002. **Anais...** Association for Computational Linguistics, 2002. p.79-86.

RUSSEL, S.; NORVIG, P. **Inteligência Artificial**. Rio de Janeiro: Elsevier Editora, 2013.

SAVADOVSKY, P. **Introdução ao Projeto de Interfaces em Linguagem Natural**. São Paulo: SID Informática, 1988.