Comparação de técnicas de mineração de opinião para identificação de emoções em tweets

GUSTAVO DE MACEDO PINTO (IFPB, Campus Picuí), LUCAS EDUARDO QUEIROZ DANTAS (IFPB, Campus Picuí), ALEXANDRE SOUTO MEDEIROS (IFPB, Campus Picuí), ANDRÉ LUIZ FIRMINO ALVES (IFPB, Campus Picuí).

E-mails: gmacedo.pinto@gmail.com, lucaseduardomdg@gmail.com, alexandre.soutomds@gmail.com, andre.alves@ifpb.edu.br.

Área de conhecimento:(Tabela CNPq): 1.03.03.04-9 Sistemas de Informação.

Palavras-Chave: processamento de linguagem natural; análise de sentimentos; machine Learning;

1. Introdução

A crescente interatividade entre os serviços oferecidos na Web e os seus usuários geram uma enorme quantidade de informação. Com esta nova forma de usar a Web, chamada de Web 2.0, os usuários não navegam simplesmente na Web, eles contribuem ativamente com o seu conteúdo por meio das aplicações, colaborando assim para a formação de uma inteligência coletiva (O'REILLY, 2007). Essa inteligência coletiva se espalhou para diversas áreas, especialmente nas relacionadas com a vida cotidiana, tais como comércio, turismo, educação e saúde, fazendo com que a Web Social se expanda exponencialmente (FELDMAN, 2013). Deste modo, compreender o que as pessoas estão pensando ou suas opiniões é fundamental para a tomada de decisões de grandes corporações, principalmente neste contexto em que as pessoas expressam seus comentários de forma voluntária no intuito de cooperar umas com as outras.

Com um alto fluxo de dados relacionados a opiniões, as empresas estão cada vez mais interessadas em obtê-las, pois essas informações podem melhorar as relações entre empresa e cliente, aprimorando a satisfação do cliente e fazendo que o consumidor sempre procure o seu serviço/produto. Porém, analisar a enorme quantidade de texto não estruturado produzido nas diversas mídias sociais é uma tarefa árdua para os humanos (KINTO; HERNANDEZ, 2016).

Desta forma, o objetivo deste trabalho é comparar técnicas de Processamento de Linguagem Natural que permitem a identificação de textos que contém emoções ou ironias. Os textos utilizados neste trabalho são *tweets* relacionados aos cortes de verbas da educação ocorridos no Brasil em 2019. A melhor técnica implementada apresenta uma acurácia de 87%. O resultado desta pesquisa pode contribuir nas tomadas de decisões do governo Brasileiro, coletando as opiniões e emoções da população sobre a educação, entre outros diversos temas.

2. Materiais e Métodos

A Figura 1 apresenta as etapas da metodologia adotada neste trabalho.

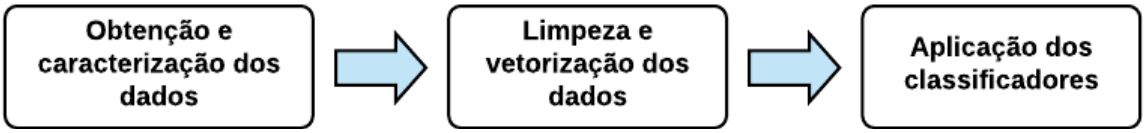

Figura 1: Etapas de desenvolvimento do trabalho.

O conjunto de dados contém mais de 70.000 *tweets e* foi coletado por Lima et al. (2020). Os dados são relacionados à temática do corte de verbas da educação que ocorreram no ano de 2019. Foram selecionados de forma aleatória cerca de 2500 *tweets* e disponibilizado em uma plataforma colaborativa Web com o objetivo de voluntários realizarem anotações de emoções (raiva, amor, alegria, tristeza, medo, ironia) presentes nos textos ou não, gerando assim um conjunto de dados anotados para aplicação das técnicas de aprendizagem de máquina.

Considerando os dados rotulados, os textos foram classificados em três categorias: Sem Emoção, Com Emoção (raiva, amor, alegria, tristeza ou medo) e com Ironia. O objetivo é possibilitar o treinamento de um algoritmo que identifique textos que contém emoções ou ironias. A Figura 2 apresenta a sumarização dos tweets anotados pelos voluntários.

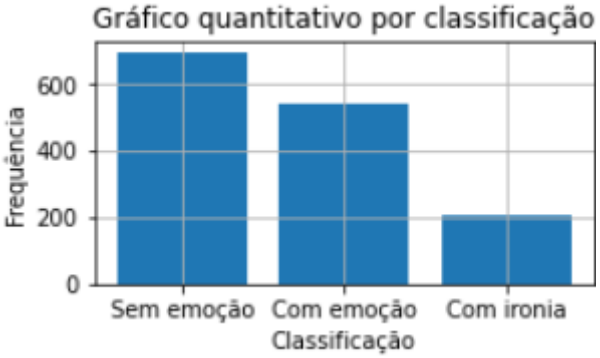

Figura 2: Tweets anotados pelos voluntários.

Os dados coletados passaram por uma etapa de pré-processamento, na qual foi realizada limpeza de ruídos, conversão de todas as palavras dos textos em minúsculas, remoção de links, acentos, linhas em branco e caracteres especiais, exceto as *hashtags*, pois as *hashtags* atribuem sentido ao texto. Também foram retiradas as *StopWords* (preposições, artigos e palavras que são consideradas irrelevantes para o modelo), pois não interferem nos modelos de classificação de textos. Esse pré-processamento foi realizado utilizando as biblioteca *nltk (Natural Language Toolkit)* da linguagem de programação *Python*, que é adequada para tratar dados de linguagem natural.

A vetorização é uma etapa extremamente necessária para a permitir a entrada de dados dos algoritmos de classificação. Na vetorização os textos são transformados em vetores numéricos de forma automática. Neste trabalho foi utilizado o *Bag of Words* que consiste em atribuir um número a palavra de acordo com a ordem que ela aparece, e computa a quantidade de repetições da palavra no texto, deste modo determinando a relevância da palavra no texto.

Na etapa de classificação de textos foram utilizadas as seguintes técnicas de *Machine Learning: Support Vector Machine (SVM)*, *Naive Bayes* e Regressão Logística. As técnicas foram avaliadas utilizando o processo de validação cruzada (***k-fold***), Esse método consiste em dividir aleatoriamente o conjunto total de dados em *k* subconjuntos mutuamente exclusivos do mesmo tamanho, onde foi separado em 3 subconjuntos. A partir disso, um subconjunto é utilizado para teste e os *k-1* restantes são utilizados para treinar o modelo.

Dos tweets anotados, 60% foi dedicado ao treino, e 40% para a validação e teste. A comparação entre as técnicas foi realizada utilizando as métricas de avaliação *Accuracy, Precision, Recall* e *F1-Score*. Essas métricas foram utilizadas nas etapas de treinamento, validação e teste dos algoritmos avaliados neste trabalho.

3. Resultados e Discussão

A Tabela 1 apresenta as métricas de avaliação utilizadas no treinamento dos classificadores implementados, possibilitando a comparação dos resultados dos algoritmos.

Classificador	Accuracy	Precision	Recall	F1-score
SVM	0.876	0.879	0.876	0.874
Naive Bayes	0.751	0.758	0.751	0.748
Regressão Logística	0.846	0.854	0.846	0.841

Tabela 1: Métricas de avaliação de desempenho.

De acordo com os resultados da Tabela 1, observa-se que o algoritmo *SVM* teve um melhor desempenho em comparação com os demais, obtendo 87,6% de Acurácia e 87,4% de *F1-Score*. A Figura 3 apresenta o desempenho do algoritmo *SVM,* cujo obteve o melhor resultado e foi submetido a validação cruzada (*Cross-Validation Score*).

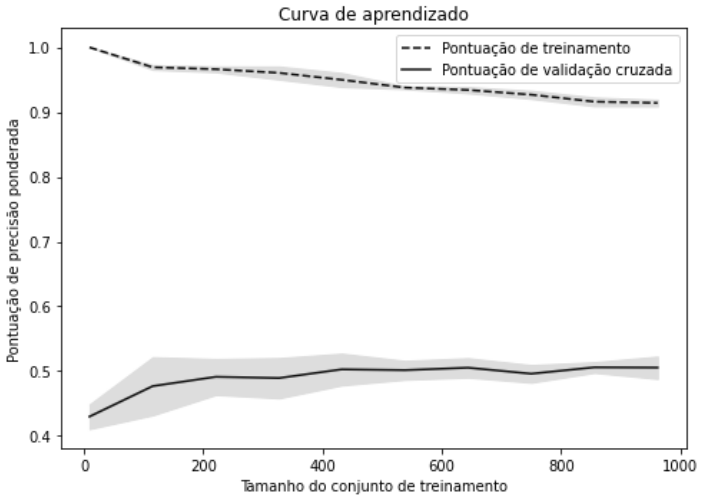

Figura 3: Cross-Validation Score: SVM.

Os resultados da validação cruzada (*Cross-Validation*) indicam *overfitting* no processo de aprendizagem dos algoritmos, pois a porcentagem de acertos no processo de treinamento diverge muito da validação-cruzada, indicando que os algoritmos estão se ajustando demais aos dados do treinamento, sendo necessário refinar os parâmetros das técnicas utilizadas e analisar o conjunto de dados do treinamento.

4. Considerações Finais

Este trabalho apresentou uma comparação entre técnicas de Processamento de Linguagem Natural para obter um classificador de textos que detecta textos com emoções e ironias. Textos de mídias sociais relacionados aos cortes de verba da educação foram coletados e rotulados de forma colaborativa com emoções para serem utilizados no processo de treinamento e avaliação das técnicas. Os classificadores comparados neste trabalho foram implementados utilizando as técnicas de *SVM*, *Naive Bayes* e Regressão Logística. Durante o processo de treinamento, o *SVM* foi a técnica que obteve o melhor resultado. No entanto, analisando com mais detalhes, foi observado que os algoritmos apresentaram *overfitting* no processo de aprendizado, sendo necessário em futuros trabalhos analisar e estudar técnicas de reamostragem *(oversampling)* para balancear os dados de treinamento e aplicar outros algoritmos de aprendizagem de máquina, como *deep learning* sendo estes trabalhos futuros. Deseja-se também, após validação das técnicas, classificar de forma automática todos os demais textos coletados, possibilitando assim uma análise sumarizada das opiniões e emoções da população em relação à temática.

Agradecimentos

Os autores agradecem ao CNPq/PIBIC-EM e IFPB por financiarem esta pesquisa.

Referências

LIMA et al. Análise de Sentimentos em Tweets: um Estudo de Caso sobre os cortes de orçamentos nas IFEs. ENCONTRO NACIONAL DE COMPUTAÇÃO DOS INSTITUTOS FEDERAIS (ENCOMPIF), 7. , 2020, Disponível em: <https://sol.sbc.org.br/index.php/encompif/article/view/11064>. Acesso em: 22 de jul. 2021.

FELDMAN, Ronen. Techniques and applications for sentiment analysis. Communications of the ACM, vol 56, Issue 4, p82, 2013. Disponível em: <https://cacm.acm.org/magazines/2013/4/162501-techniques-and-applications-for-sentiment-analysis/fulltext>. Acesso em: 25 de jul. 2021.

O'REILLY, Tim. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. Communications & Strategies, Vol 1, Issue 65, Pages 17-37, 2007. Disponível em: <https://mpra.ub.uni-muenchen.de/id/eprint/4580>. Acesso em: 19 de jul. 2021.

KINTO, Eduardo; HERNANDEZ, Emílio. Classificadores de Texto Reduzido Baseados em SVM. Sociedade Brasileira de Inteligência Computacional (SBIC), São Paulo, 2016. Disponível em: <http://abricom.org.br/eventos/cbrn_2005/cbrn2005_096/>. Acesso em: 28 de jul. 2021.