TÉCNICAS DE PRÉ-PROCESSAMENTO DE TEXTOS ADVINDOS DA WEB

RODRIGO BARBOSA CARVALHO[1]; LARISSA ASTROGILDO DE FREITAS[2]

[1]*Universidade Federal de Pelotas – rbcarvalho@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – larissa@inf.ufpel.edu.br*

1. INTRODUÇÃO

Atualmente é possível acessar uma enorme quantidade de informação em diversos locais da internet (como blogs, redes sociais, sites de entretenimento, sites de compra ou venda de produtos, etc.), informação esta que pode ser obtida de maneira rápida e de diversos dispositivos de fácil acesso.

Para a Análise de Sentimento (França, Oliveira, 2014), área de pesquisa destinada à processar e identificar sentimentos e emoções em sentenças da língua natural, esse é um fator muito positivo. Porém não é recomendado processar os dados de forma crua pois, para atingir uma interpretação de alta qualidade, é necessário utilizar do pré-processamento adequado para cada caso, analisando componentes presentes na informação e o domínio da mesma (Gurudath, 2020).

Partindo deste princípio, este documento tem como objetivo apresentar algumas das formas de pré-processamento existentes para o tratamento de informação vinda de alguns contextos presentes na internet, explicando seus conceitos e funcionalidades.

2. METODOLOGIA

Enquanto usamos o termo “pré-processamento”, estamos nos referindo à todo e qualquer tratamento da informação feito antes de realizar de fato qualquer análise dos dados (Ujjawal, 2020). Normalmente é associado à esse termo operações como: remoção de espaços em branco extras, expansão de contrações das palavras, tornar todos os caracteres minúsculos, categorizar as palavras com *POS* (*Part of Speech,* ou parte do discurso, operação que consiste em classificar todas as palavras da sentença de acordo com a sua classe gramatical e quais palavras estão relacionadas no texto), reconhecimento de entidades nomeadas, entre outras possíveis técnicas.

Contudo, durante os experimentos, iremos abordar técnicas mais amplas e específicas do domínio da informação utilizada. Ao longo dos experimentos usaremos de dois *datasets* os quais consistem de *tweets*, o primeiro intitulado como *TweetSentBR*[1], *dataset* o qual contém 15000 *tweets* de contexto geral retirados da plataforma *Twitter*[2], e o segundo sendo um *dataset* encontrado na plataforma Kaggle[3], o qual consiste em 8199 *tweets* referentes à assuntos gerais que remetem ao estado de Minas Gerais, com uma presença significativa de assuntos policiais.

Como mencionado no parágrafo anterior, utilizamos métodos mais precisos do domínio da informação, mais especificamente, procedimentos que são aproveitáveis somente neste tipo de informação. Para o experimento, como a plataforma *Twitter* permite que pessoas repliquem postagens realizadas por

[1] https://sites.google.com/icmc.usp.br/opinando/
[2] https://twitter.com/
[3] https://github.com/minerandodados/mdrepo/blob/master/Tweets_Mg.csv

outros, iremos admitir que estas informações não são genuínas, assim o primeiro pré-processamento será a remoção dos *retweets* presentes nos *datasets*.

Em um segundo momento, foi analisado uma forma de apresentação de tópicos nas sentenças utilizadas, denominada como *hashtags*. Esta forma de apresentação de conteúdo é usada para estabelecer uma referência à algum assunto pertinente à outro contexto, portanto como trata-se apenas de uma forma de ligar a sentença à outra, iremos considerar durante os experimentos que elas são desprovidas do sentimento gerado pelo autor, assim sendo desconsideradas e extraídas do texto.

Da mesma forma, estão presentes links para outros sites inseridos no conteúdo e, por ser semelhante ao uso das *hashtags* descrito acima, serão considerados como semelhantes e também serão retirados da informação. Com essas considerações, esta será a outra forma de pré-processamento a ser utilizada.

Para executar a tarefa de avaliar o sentimento contido nas sentenças escolhidas, foi utilizado como opção de léxico o LeIA (Léxico para Inferência Adaptada), um fork do léxico e da ferramenta *VADER* (*Valence Aware Dictionary and sEntiment Reasoner*), presente na biblioteca Python NLTK (*Natural Language ToolKit)*, adaptado para ser utilizado no português brasileiro e focado em análises de textos advindos de redes sociais (Almeida, 2018).

3. RESULTADOS E DISCUSSÃO

Como critério de avaliação utilizamos: a métrica acurácia, que consiste na porcentagem das verificações que foram realizadas corretamente, em relação à todas as polaridades possíveis; a métrica precisão, que representa quantos acertos foram realizados pelo léxico para cada classificação possível; e a métrica *recall*, que representa quantas das sentenças presentes no *dataset* foram devidamente avaliadas, para cada polaridade. No caso deste experimento, as classificações do léxico e do *dataset* são “positivo” (Pos.), “negativo” (Neg.) e “neutro” (Neu.), pois ambos operam em um sistema de 3 classes. Porém devemos também analisar os efeitos que foram produzidos nos dados após os pré-processamentos serem aplicados.

Após a realização dos experimentos propostos, adquiriu-se os seguintes resultados, como mostram as tabelas:

Dataset		Precisão	Recall	Acurácia
Kaggle	Neg.	0,24	0,35	29,34
	Neu.	0,33	0,42	
	Pos.	0,33	0,16	
TweetSentBR	Neg.	0,50	0,49	47,65
	Neu.	0,34	0,51	
	Pos.	0,64	0,45	

Tabela 1: Valores Sem Pré-processamentos

Na Tabela 1 vemos os valores originais dos *datasets* ao serem aplicados no LeIA, a opção de léxico escolhida para a realização dos testes..

Dataset		Precisão	Recall	Acurácia
Kaggle	Neg.	0,08	0,24	25,26
	Neu.	0,38	0,44	
	Pos.	0,37	0,14	
TweetSentBR	Neg.	0,50	0,49	47,65
	Neu.	0,34	0,51	
	Pos.	0,64	0,45	

Tabela 2: Valores Com Pré-processamentos

Como observado nos resultados demonstrados pela Tabela 2, os métodos aplicados surtiram efeito nas características do teste realizado com o *dataset* da plataforma Kaggle, tendo variações tanto positivas quanto negativas em seus valores, com uma aparente redução da sua acurácia, porém, ao aplicar o pré-processamento nos seus dados, a consistência dos mesmos foi significantemente aumentada, a ponto de retirar informação desnecessária referente à mais de um terço da informação total, tornando qualquer experimento realizado com ele mais íntegro e confiável.

Por outro lado, não houve nenhuma alteração no teste realizado com o *dataset TweetSentBR*, isso se dá ao fato de que este conjunto de dados contém uma qualidade consistente com dados que já sofreram um ou mais pré-processamentos, assim as alterações realizadas não surtiram efeito e mantiveram os valores de ambos os testes iguais.

4. CONCLUSÕES

Relevando os efeitos causados nos dados utilizados, é justo afirmar que os métodos e a quantidade de pré-processamento que um conjunto de dados requer é completamente diferente de um caso para outro, assim sendo necessária uma avaliação da qualidade e integridade do estado em que os dados se encontram para uma boa tomada de decisão, já que, como demonstrado pelo experimento, o pré-processamento de informação é uma ferramenta essencial no aprimoramento da integridade de qualquer experimento.

Partindo desta colocação, uma provável continuação deste experimento seria a inversão das relevâncias consideradas, isto é, ao invés de desconsiderar *hashtags* e *links*, modelá-los em alguma forma de considerar a informação neles contida. Outra abordagem seria alterar o domínio da informação, saindo de conjuntos de dados vindos de redes sociais e partindo para *datasets* de portais de compra e venda, *reviews* de filmes e livros, entre tantos outros.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Tiago França and Jonice Oliveira. 2014. Análise de Sentimento de Tweets Relacionados aos Protestos que ocorreram no Brasil entre Junho e Agosto de 2013. In III Brazilian Workshop on Social Network Analysis and Mining. SBC, 128–139.

Gurudath, S. Natural Language Processing. Acessado em 25 set. 2020. Online Disponível em: https://medium.com/analytics-vidhya/natural-language-processing-bedb2e1c8ceb

Ujjawal, V. Text Processing for NLP (Natural Language Processing),Begginers to Master. Online. Disponível em: https://medium.com/analytics-vidhya/text-preprocessing-for-nlp-natural-language-processing-beginners-to-master-fd82dfecf95

Rafael J. A. Almeida. 2018. LeIA - Léxico para Inferência Adaptada. Acessado em 25 set. 2020. Disponível em: https://github.com/rafjaa/LeIA.