Análise de sentimento em português

Aluna: Elaine Costa da Cunha.

Orientador: Leonardo Forero Mendoza.

Análise de sentimento em português

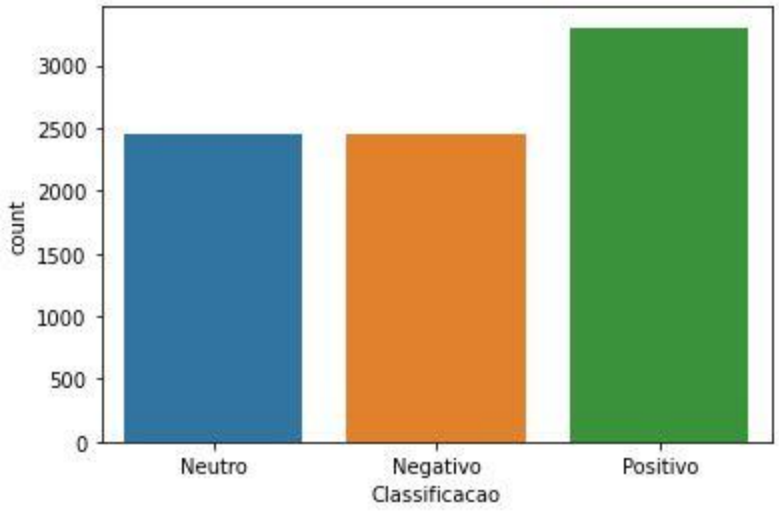
Trabalho apresentado ao curso BI MASTER como pré-requisito para conclusão de curso e obtenção de crédito na disciplina "Projetos de Sistemas Inteligentes de Apoio à Decisão".

Resumo

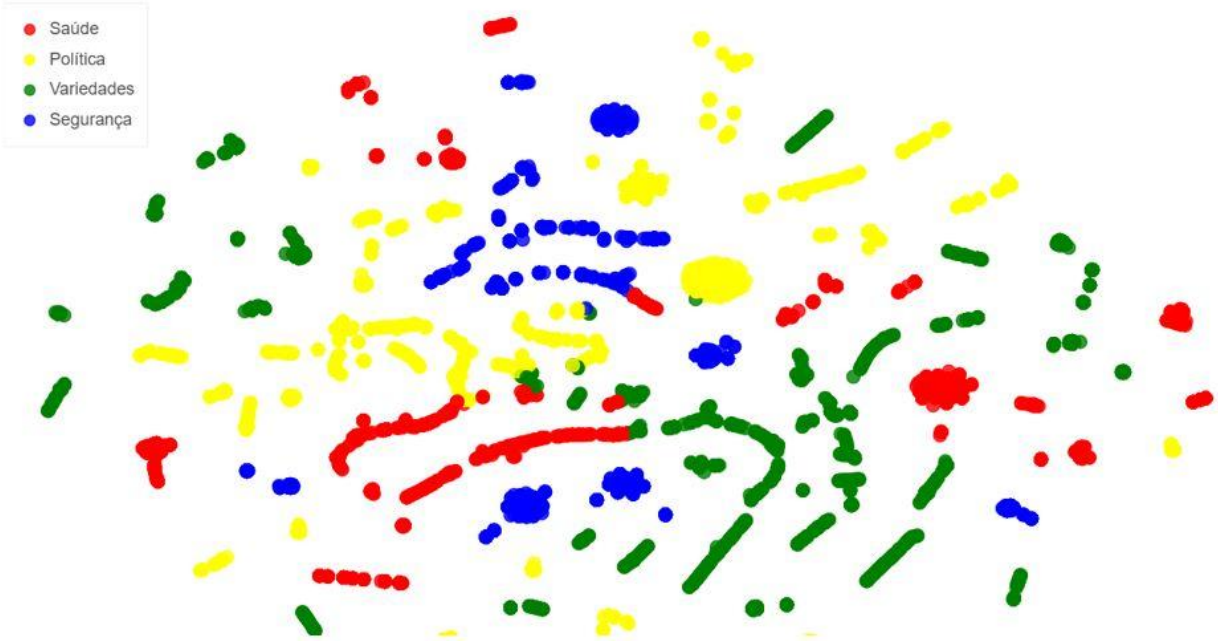
A literatura tem uma maior abordagem do assunto Análise de Sentimento no idioma em inglês. O desafio deste trabalho consiste em fazer testes em uma base do Twitter no idioma português, usando uma modelagem com e sem o uso de embeddings aplicados a dificuldades relacionadas às acentuações e nuances do próprio idioma. Não existe um resultado definitivo, foram aplicados três códigos com abordagens um pouco diferentes. Cada variação no código como a forma de limpeza da base, uso de tokenização, lemetização e outros filtros, redes neurais, tudo faz parte do teste. A análise da base antes e depois de pré-processamento, uso de deep learning, modelos estatísticos trouxeram uma abrangência de conhecimento dos recursos que a análise de sentimento pode trazer para diversas áreas de interesse.

Base de dados

A base em português utilizada veio do Kaggle (https://www.kaggle.com/faelk8/portuguese-sentiment-analysis). Esta é uma base de tweets que tem 8199 linhas e aborda diversos assuntos.A quantidade de classificações como positivo, neutro, negativo serviu também para análise de comportamento de determinado tipo classificação.



Com uso de representações por frequência foi possível identificar algumas palavras mais recorrentes que se encaixam em determinados assuntos.



Análise do dataset e pré-processamento

Na etapa de limpeza de texto o algoritmo que usou LDA teve uma melhora bem expressiva na distinção entre os tópicos. Isto reforça o quanto esta parte de pré-processamento, a que exigiu maior tempo da POC, é uma das mais importantes para o resultado deste trabalho.

A limpeza da base consistiu em retirada de símbolos, urls, hashtags, re-tweets, caracteres especiais, mas como o texto é da língua portuguesa a parte de acentuação foi mantida. O texto foi colocado em letras minúsculas.

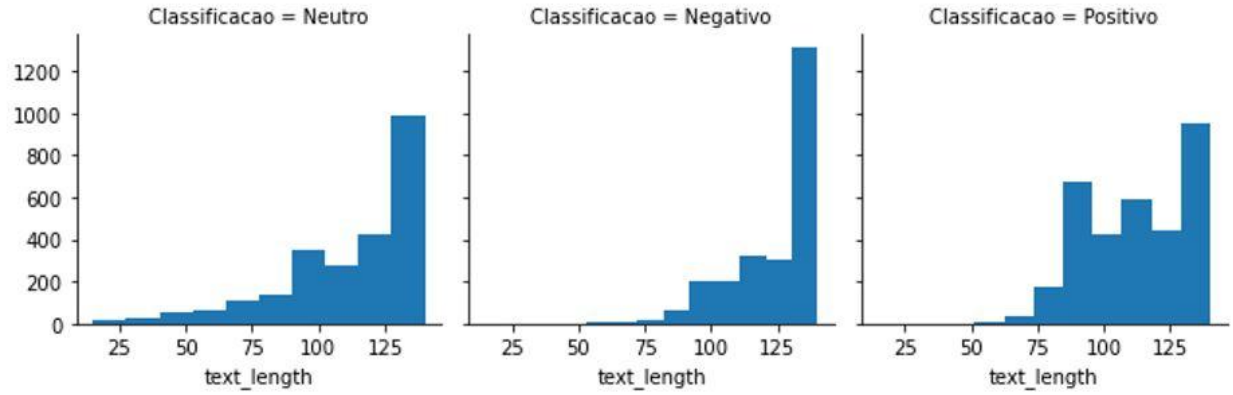
Na elaboração da nuvem de palavras foram aplicados os n-grams. Os n-grams são junções de palavras, em algumas situações a junção de 2, 3 palavras trouxe mais sentido a um determinado contexto.

Unigram

Bigrams

Trigrams

Foram avaliados tamanho das palavras e caracteres por quantidade de labels (positivo, negativo e neutro).

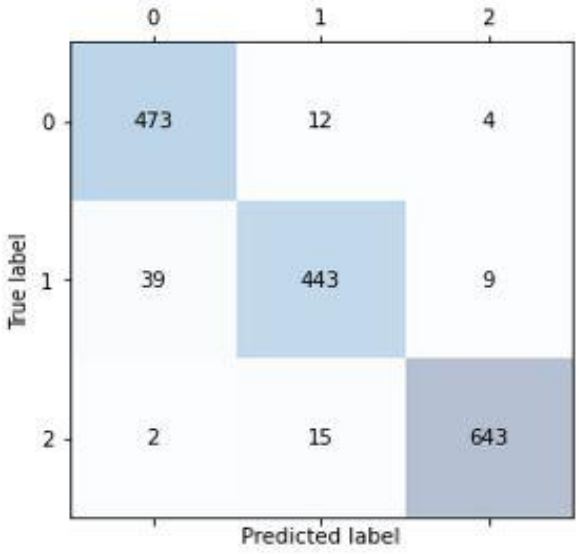


Redes Neurais

No uso da rede neural CNN-LSTM, os parâmetros foram configurados de forma semelhante, diferenciando número de camadas, e a aplicação do uso de embeddings.

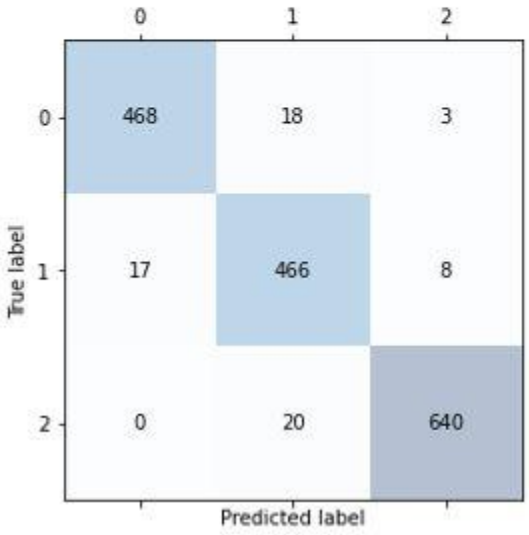
Sem embeddings

	precision	recall	f1-score	support
0	0.92	0.97	0.94	489
1	0.94	0.90	0.92	491
2	0.98	0.97	0.98	660
accuracy			0.95	1640
macro avg	0.95	0.95	0.95	1640
weighted avg	0.95	0.95	0.95	1640



Com embeddings

	precision	recall	f1-score	support
0	0.96	0.96	0.96	489
1	0.92	0.95	0.94	491
2	0.98	0.97	0.98	660
accuracy			0.96	1640
macro avg	0.96	0.96	0.96	1640
weighted avg	0.96	0.96	0.96	1640



Conclusão

O objetivo deste trabalho está em testar diferentes formas de avaliar a análise de sentimento através de algoritmos de deep learning e aplicação de modelos estatísticos. Os resultados de acurácia ficaram próximos com o uso da rede neural com e sem embeddings em torno de 95% e 96%.

Considerações finais

Gostaria de agradecer a todos os professores do curso, em especial ao professor Leonardo Mendoza pelos ensinamentos e orientação neste trabalho.

Matrícula: 192.671.087

Pontifícia Universidade Católica do Rio de Janeiro
Curso de Pós Graduação *Business Intelligence Master*