PROCESSAMENTO DE LINGUAGEM NATURAL - UMA ANÁLISE DAS NOTÍCIAS RELACIONADAS AO SARS-COV-2

Hugo T. M. Oliveira
Instituto de Computação
Universidade Federal de Alagoas
Maceió, Alagoas
htmo@ic.ufal.br

Valério N. R. Júnior
Instituto de Computação
Universidade Federal de Alagoas
Maceió, Alagoas
vnrj@ic.ufal.br

28 de Janeiro de 2022

ABSTRACT

No contexto da pandemia de COVID-19 a imprensa tem desempenhado um papel ainda mais importante na sociedade. Na era da informação, o processamento de linguagem natural pode ser utilizado para minerar informação dos textos. Neste trabalho, forma utilizadas técnicas de pré-processamento no corpus para realizar tarefas como classificação, agrupamento, visualização e análise sintática, aplicando abordagens bem estabelecidas e comparando-as. Como resultado, temos desde a classificação de notícias entre dois meios de comunicação distintos à extração de informações sobre pessoas presentes nas notícias.

1 Introdução

Diante dos desafios impostos pela pandemia da COVID-19 (Sars-CoV-2) no mundo, é natural que surjam dúvidas em relação às notícias que tratam do tema, publicadas nos meios de comunicação, uma vez que o volume de informação que chega ao ouvinte final é avassalador. É ainda mais relevante, tentar entender quais os principais assuntos abordados quando o tema que se discute é o novo coronavírus ou, por exemplo, quais as principais entidades (pessoas, instituições ou organizações) envolvidas nas principais manchetes.

É nesse sentido que buscamos propor um trabalho onde coletamos diversos artigos de notícias relacionados pelo termo *coronavírus* disponíveis em dois dos principais portais de notícia do Brasil: o portal da CNN Brasil e BBC Brasil. O objetivo é justamente tentar responder às questões formuladas anteriormente ou até mesmo traçar um "perfil de escrita"de cada portal, propondo modelos de aprendizagem de máquina que classifiquem corretamente artigos de redações diferentes.

Um fato importante a ressaltar é a utilização ampla de conceitos e técnicas relacionados ao processamento de linguagem natural para propor modelos que trabalhem com os artigos de notícias, uma vez que os textos são escritos de modo que seres humanos entendam, e não máquinas. Além disso, dificuldades como a própria formatação das páginas *web* tiveram que ser contornadas, para que pudéssemos separar o texto útil da notícia do restante do conteúdo irrelevante da página.

Diversos experimentos foram realizados nos textos: a classificação de documentos, o reconhecimento de entidades nomeadas e a modelagem de tópicos. Os resultados obtidos sugerem que de fato, sob o ponto de vista dos documentos coletados no período de Março à Junho de 2020 as redações davam preferência a abordar temas sutilmente distintos.

1.1 Trabalhos Relacionados

Estudar o papel de atuação dos veículos de imprensa na divulgação de notícias relacionadas a crises sanitárias sempre foi um tema abordado por pesquisadores uma vez que diversas políticas públicas podem ser afetadas pela maneira que são veiculadas nos meios de comunicação. Dessa forma, trabalhos foram propostos de maneira a entender

como as notícias divulgadas pelos principais meios de comunicação podem afetar o comportamento das autoridades governamentais, como por exemplo em [1].

Tendo em vista que o tema escolhido é bastante recente, uma vez que ainda estamos enfrentando a pandemia globalmente, os trabalhos que investigam a crise da pandemia da COVID-19 e o que está sendo publicado nas notícias envolvem abordagens diversas que surgem a medida que vão se colhendo mais informações acerca do que se é transmitido pelos veículos. Um trabalho relacionado que podemos citar foi proposto em [2], coletando manchetes de notícias relacionadas ao termo *coronavirus* utilizando, por exemplo, técnicas de processamento de linguagem natural como análise de sentimento para classificar cada manchete em uma "emoção"específica ou criando nuvem de palavras com termos mais frequentes nas manchetes.

2 Metodologia

Os experimentos propostos foram realizados numa base de textos, construída através da consulta por artigos de notícias relacionados ao novo coronavírus. As consultas foram feitas através de ferramentas de *web scraping* disponíveis nas linguagens de programação *Javascript* e *Python*. Com o *corpus* estabelecido, os textos passaram por uma etapa de limpeza e pré-processamento antes que fossem utilizados em cada experimento.

Todas as ferramentas utilizadas para implementar os modelos e processar o textos vieram da linguagem de programação *Python* que conta com um acervo extensivo de pacotes e *frameworks* para a rápida implementação de modelos de Aprendizagem de Máquina e Processamento de Linguagem Natural.

A maioria dos experimentos foi executada no Google Colab[3] e Jupyter Notebook[4], aplicações *web* que permitiram a edição e compartilhamento de documentos interativos com o código-fonte em *Python* dos algoritmos implementados para cada abordagem proposta.

2.1 Descrição do corpus

O corpus utilizado neste trabalho consiste em notícias em português relacionadas ao novo coronavírus (Sars-CoV-2) extraídas dos sites http://www.bbc.com.br e http://www.cnn.com.br. Os artigos de notícias foram coletados através de buscas em cada site (*web scraping*) pela palavra chave *coronavírus*.

No total há 1625 textos, sendo 350 da BBC e 1275 da CNN. Para cada página encontrada através da busca foi extraído o arquivo fonte da página (HTML). Com base nas *tags* de marcação dos elementos de cada página o arquivo *.html* foi processado para formar um corpo único de texto composto por um título principal e uma coleção de parágrafos, separando o conteúdo das notícias do restante da página web.

2.2 Pré-processamento

A primeira etapa do pré-processamento consistiu na *higienização* dos dados. Nessa etapa, foram aplicadas as seguintes modificações no texto bruto:

- Substituição das quebras de linha por espaços
- Remoção de fragmentos de texto que se repetem na maioria dos documentos, sem algum significado especial, como parte do cabeçalho da página)
- Remoção de todas as palavras que contém números
- Remoção da pontuação e conversão para caixa baixa

Com o texto higienizado, para os problemas de classificação de documentos, modelagem de tópicos, agrupamento e visualização com projeção multidimensional o pré-processamento realizado consistiu na remoção dos termos "bbc"e "cnn", remoção de *stopwords* e por fim *stemmização*.

2.3 Abordagens e técnicas adotadas

Para resolver os problemas propostos foram utilizadas várias técnicas de aprendizado de máquina e álgebra linear. Em alguns problemas, como a classificação de documentos, diferentes abordagens foram testadas, indo de modelos clássicos como Support Vector Machine[5] (SVM) até modelos tidos como estado da arte através de aprendizagem profunda. Nesta seção serão descritas as técnicas utilizadas no contexto do problema envolvido.

2.3.1 Classificação de documentos

A tarefa de classificação de documentos foi um dos experimentos propostos com a base de textos, onde a abordagem utilizada foi baseada em mais de um modelo: através de métodos de extração de *features* e classificação clássicos da aprendizagem de máquina (SVM, *Naive Bayes*[6] e Regressão Logística) e métodos mais sofisticados de aprendizagem profunda: Redes Neurais Recorrentes *Long Short-Term Memory[7] (LSTM)*.

Os modelos foram treinados de forma supervisionada para classificação binária: artigos de notícia rotulados como BBC e CNN. Para balancear o conjunto de dados, limitamos a quantidade de artigos pelo número de exemplos da menor classe (350 documentos do tipo BBC) totalizando um conjunto com 700 textos rotulados.

Na abordagem clássica, utilizamos como *feature extraction* os métodos *Count Vectorization* e *Term Frequency–Inverse Document Frequency[8] (TfIdf)* que se baseiam na ocorrência das palavras do vocabulário em cada texto. Nesse sentido, cada documento foi mapeado para um vetor do $\mathbb{R}^V$ onde $V \sim 3 \times 10^4$. Com isso, pudemos treinar os modelos de classificação e, adicionalmente, determinar medidas de similaridade entre os textos.

Na abordagem envolvendo *LSTMs*, experimentamos diferentes arquiteturas de rede para determinar a que melhor desempenhava a classificação corretamente (acurácia de validação), escolhendo diferentes parâmetros para o modelo, como será detalhado mais adiante.

2.3.2 Modelagem de tópicos

A modelagem de tópicos consiste em agrupar documentos que tratam do mesmo tema através da ocorrência de certos grupos de palavras. As técnicas utilizadas foram *Latent Dirichlet Allocation[9] (LDA)*, *Singular Value Decomposition[10] (SVD)* e *Non-negative Matrix Factorization[10] (NMF)*.

2.3.3 Agrupamento

Com base na vetorização obtida de cada documento (tal qual foi descrita na seção **2.3.1**) realizamos um agrupamento do conjunto de documentos. Primeiramente, utilizamos o algoritmo *Principal Component Analysis[**?**] (PCA)* para reduzir a dimensionalidade dos dados, preservando sua variância em 95%. Em seguida, com os pontos projetados num espaço de menor dimensão, utilizamos o algoritmo *K-means*[**?**] com o método *elbow*[**?**] para encontrar um valor de k ótimo.

2.3.4 Visualização com projeção multidimensional

A partir do agrupamento realizado na etapa anterior, prosseguimos com a visualização dos dados projetados. Utilizamos duas técnicas diferentes, *t-Distributed Stochastic Nighbor (TSNE)[11]* e *Uniform Manifold Approximation and Projection (UMAP)[12]*, para reduzir ainda mais a dimensionalidade do problema, projetando os pontos num espaço de visualização de dimensão 2. Em seguida, colorimos cada ponto do plano de acordo com o grupo determinado pelo algoritmo de agrupamento utilizado, fornecendo um gráfico de dispersão com *clusters* de pontos indicando documentos relacionados à temas similares.

Para efeito de comparação, o mesmo processo foi aplicado utilizando a representação vetorial *Doc2Vec*[**?**] de cada documento, obtendo resultados muito mais significativos de *clusterização* dos dados.

2.3.5 Extração de informação

Para extrair informação do corpus foram utilizadas as etiquetas gramaticais (*part of speech*), reconhecimento de entidades nomeadas (*NER*) e análise das dependências das árvores sintáticas.

3 Experimentos

3.1 Classificação de documentos

A primeira tarefa realizada no experimento de classificação de documentos foi tentar observar documentos "similares"de acordo com cada método de vetorização utilizado (*Count Vectorizing* e *TfIdf*). Assim utilizamos a medida de *similaridade do cosseno* entre cada vetor representante do documento para determinar o par de documentos mais similares (um para cada vetorização) na base, dado qualquer outro documento. Os resultados observados garantiram uma similaridade maior para os documentos que foram vetorizados através do *CountVectorizer*.

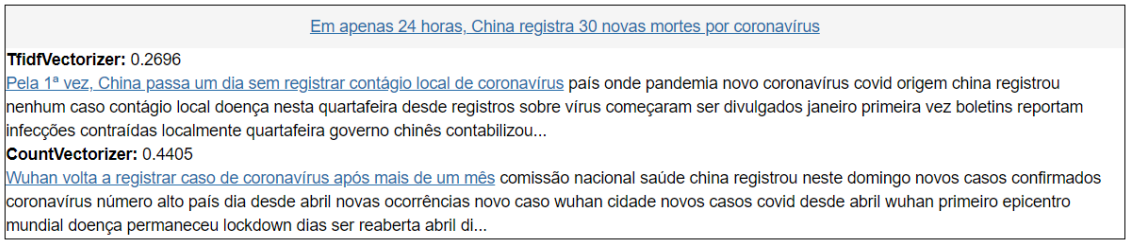


Figura 1: Par de documentos mais similares, dado um documento qualquer da base (título no topo).

Em seguida, para cada tipo de vetorização, treinamos os modelos de classificação binária obtendo os seguintes resultados (abordagem clássica):

Abordagem	Acurácia	Precisão	Cobertura	F1 Score
Regressão Logísitca + TF-IDF	0.880952	0.879630	0.887850	0.883721
Regressão Logísitca + Count Vectorizer	**0.952381**	0.939655	**0.973214**	**0.956140**
Naive Bayes + TF-IDF	0.647619	**0.967742**	0.291262	0.447761
Naive Bayes + Count Vectorizer	0.733333	0.961538	0.480769	0.641026
SVM + TF-IDF	0.866667	0.826087	0.922330	0.871560
SVM + Count Vectorizer	0.761905	0.710938	0.875000	0.784483

Tabela 1: Resultado das métricas de validação de cada abordagem (modelo + vetorização)

Posteriormente, utilizamos as redes neurais recorrentes *LSTMs* em duas principais arquiteturas, obtendo resultados superiores na classificação.

Treinando o modelo ao longo de 20 épocas com o tamanho do lote igual a 64 e separando 25% dos exemplos para validação, obtemos um resultado de **97.77%** de acurácia na validação com a seguinte arquitetura:

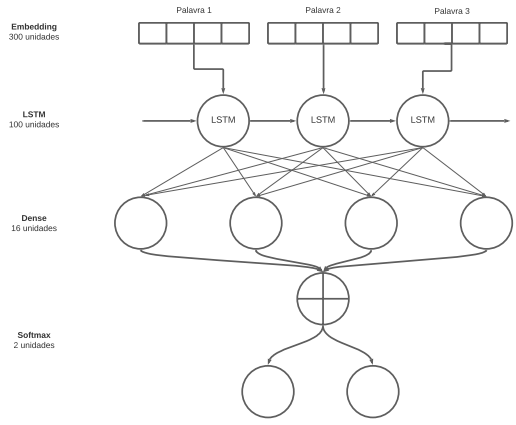


Figura 2: Arquitetura da rede com melhor desempenho

A outra arquitetura utilizada foi uma modificação da melhor, adicionando uma camada convolucional entre as camadas de *Embbeding* e *LSTM*, sem obter resultados muito diferentes (acurácia de validação próxima a **95%**).

3.2 Modelagem de tópicos

Foram identificados empiricamente 3 tópicos para os documentos. As tabelas 2, 3 e 4 comparam as 5 palavras mais relevantes em cada tópico utilizando cada abordagem (LDA, SVD e NMF), respectivamente. Diante desses resultados podemos inferir que o tópicos encontrados podem se referir a notícias sobre saúde, estatísticas da pandemia de COVID-19 e política. Das três abordagens, o melhor resultado foi obtido com a NMF.

Tópico	Palavras
1	casos, coronavírus, saúde, pessoas, brasil
2	narloch, bock, lia, molica, leandro
3	basília, assista, análise, rodrigues, junqueira

Tabela 2: Palavras mais relevantes de cada tópico com a abordagem LDA

Tópico	Palavras
1	casos, saúde, coronavírus, pessoas, brasil
2	casos, mortes, confirmados, número, rio
3	bolsonaro, presidente, ministro, federal, governo

Tabela 3: Palavras mais relevantes de cada tópico com a abordagem SVD

Tópico	Palavras
1	pessoas, vírus, diz, pode, ser
2	casos, mortes, saúde, número, confirmados
3	presidente, bolsonaro, governo, ministro, federal

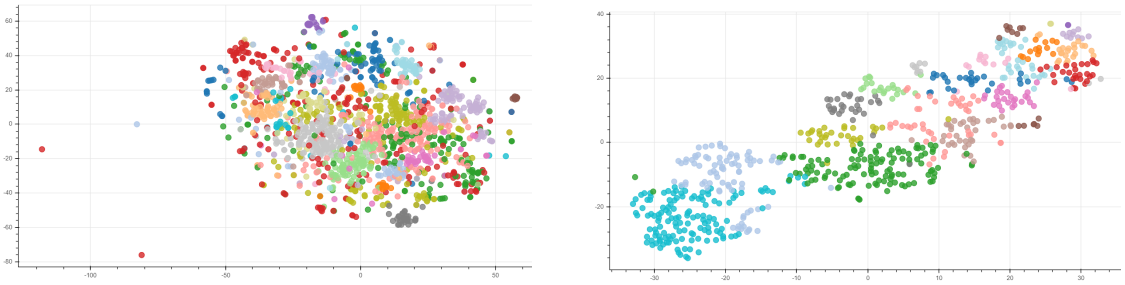
Tabela 4: Palavras mais relevantes de cada tópico com a abordagem NMF

3.3 Agrupamento e Visualização com projeção multidimensional

A redução inicial da dimensionalidade utilizando o PCA e mantendo a variância dos dados em 95% resultou num espaço de dimensão da ordem de $V' \sim 10^3$ na abordagem clássica e $V' = 4$ na representação *Doc2Vec*.

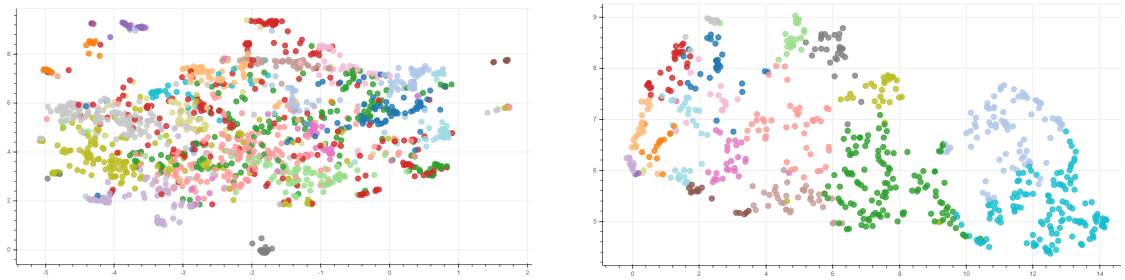
Os resultados obtidos através do algoritmo *elbow*, indicaram um valor de k próximo a 20. Com isso, utilizamos as técnicas descritas anteriormente - *TSNE* e *UMAP* - para reduzir a dimensão dos vetores para o espaço visual. A comparação de cada projeção pode ser observada nas figuras 3 e 4.

Podemos perceber a formação de grupos de documentos bem definidos utilizando a abordagem mais sofisticada *Doc2Vec*.



(a) TfIdf (b) Doc2Vec

Figura 3: Projetando os pontos utilizando *TSNE*



(a) TfIdf (b) Doc2Vec

Figura 4: Projetando os pontos utilizando *UMAP*

3.4 Extração de informação

Nesta etapa todos os experimentos foram realizados com a biblioteca *spaCy* com o conjunto de treinamento *pt_core_news_sm*. Primeiramente foram obtidas as etiquetas gramaticais das palavras. No gráfico 5 é possível visualizar como essas etiquetas estão distribuídas no corpus. Em seguida, utilizando o reconhecimento de entidades nomeadas obteve-se os nomes de pessoas. Alguns termos foram classificados de forma errada, como por exemplo "R$"e "Acho"provavelmente por iniciarem com letra maiúscula. Para compor a tabela 5 cada nome foi contado apenas uma vez por documento.

A partir da árvore sintática de cada texto foi extraída informação sobre os nomes da tabela 5 de acordo com o tipo de dependência. Na maioria das vezes, o nó raiz é um verbo e o sujeito da oração é filho desse nó. Portanto, fazendo uma busca em profundidade na árvore à procura dessas relações encontramos os resultados mostrados na tabela 6, organizados em triplas (*pessoa*, *ação*, *informação*).

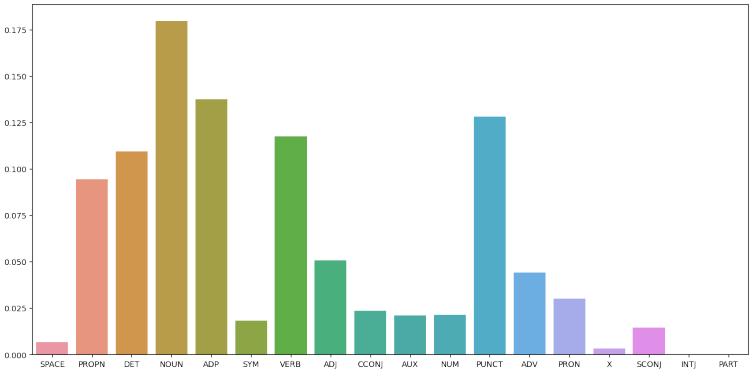


Figura 5: Distribuição de etiquetas gramaticais no corpus

Nome	Ocorrências
Tegnell	31
Trump	27
p Bolsonaro	16
Gates	15
Lotufo	15
Shiller	14
Alves	13
Chris	13
Rifkin	12
Monica	12
Rivera	11
Bremmer	11
Taísa	11
Neris	10
Bufka	10
Jesus	10
Fang Fang	10
Lincoln	10
Zimmermann	10
Witzel	9
Mandetta	9
Guedes	9
Smith	9
Atmar	9
Wilder-Smith	8
Acho	8

Tabela 5: 30 pessoas mais frequentes no corpus pela quantidade de documentos em que aparecem.

Nome	Ação	Informação
Tegnell	mantém	convicção
Tegnell	enfatizou	na
Tegnell	segue	na
Trump	criticou	esforços
Trump	apontou	cloroquina
Trump	classifica	debate
Meirelles	separar	questões
Meirelles	tem	série
Meirelles	abraça	proposta

Tabela 6: Exemplos de informações extraídas do corpus com *NER* e árvore sintática.

4 Conclusão

A partir da análise dos resultados dos experimentos, vemos que as redes *LSTM* superaram com uma ampla margem os métodos clássicos, corroborando com os avanços recentes da área de processamento de linguagem natural com aprendizagem profunda em detrimento aos métodos estatísticos. Dentre os três algoritmos usados na modelagem de tópicos o *NMF* foi o que gerou resultados mais coerentes. Quanto à projeção dos documentos, tanto o *TSNE* quanto o *UMAP* deram resultados interessantes. As duas projeções apresentaram melhoras significantes quando os vetores utilizados foram obtidos a partir do *Doc2Vec* em relação ao *Tfidf*. Já na extração de informações os resultados podem melhorar a partir de um modelo pré-treinado mais robusto.

Referências

[1] Rita de Cassia Barradas Barata. Saúde e direito à informação. *Cadernos de Saúde Pública*, pages 385–399, 1990.

[2] F. Aslam, T.M. Awan, and J.H. Syed. Sentiments and emotions evoked by news headlines of coronavirus disease (covid-19) outbreak. *Humanities and Social Sciences Communications*, 2020.

[3] Colaboratory. `http://colab.research.google.com`. Accessed: 2020-10-08.

[4] Jupyter notebook. `http://jupyter.org`. Accessed: 2020-10-08.

[5] Theodoros Evgeniou and Massimiliano Pontil. Support vector machines: Theory and applications. volume 2049, pages 249–257, 01 2001.

[6] David Hand and Keming Yu. Idiot's bayes: Not so stupid after all? *International Statistical Review*, 69:385 – 398, 05 2007.

[7] Understanding lstm networks. `https://colah.github.io/posts/2015-08-Understanding-LSTMs/`. Accessed: 2020-10-08.

[8] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[10] S. Arora, R. Ge, and A. Moitra. Learning topic models – going beyond svd. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 1–10, 2012.

[11] Geoffrey Hinton and Sam Roweis. Stochastic neighbor embedding. 15, 06 2003.

[12] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.