UMA ABORDAGEM DE CIÊNCIA DE DADOS PARA IDENTIFICAR FAKE NEWS NO ÂMBITO POLÍTICO

MARCELO HIDEAKI IWATA KITO

O QUE FOI FEITO

ATUALIZAÇÃO DOS DADOS

- Reexecução do crawler para o site **boatos.org** novamente para pegar dados mais recentes;

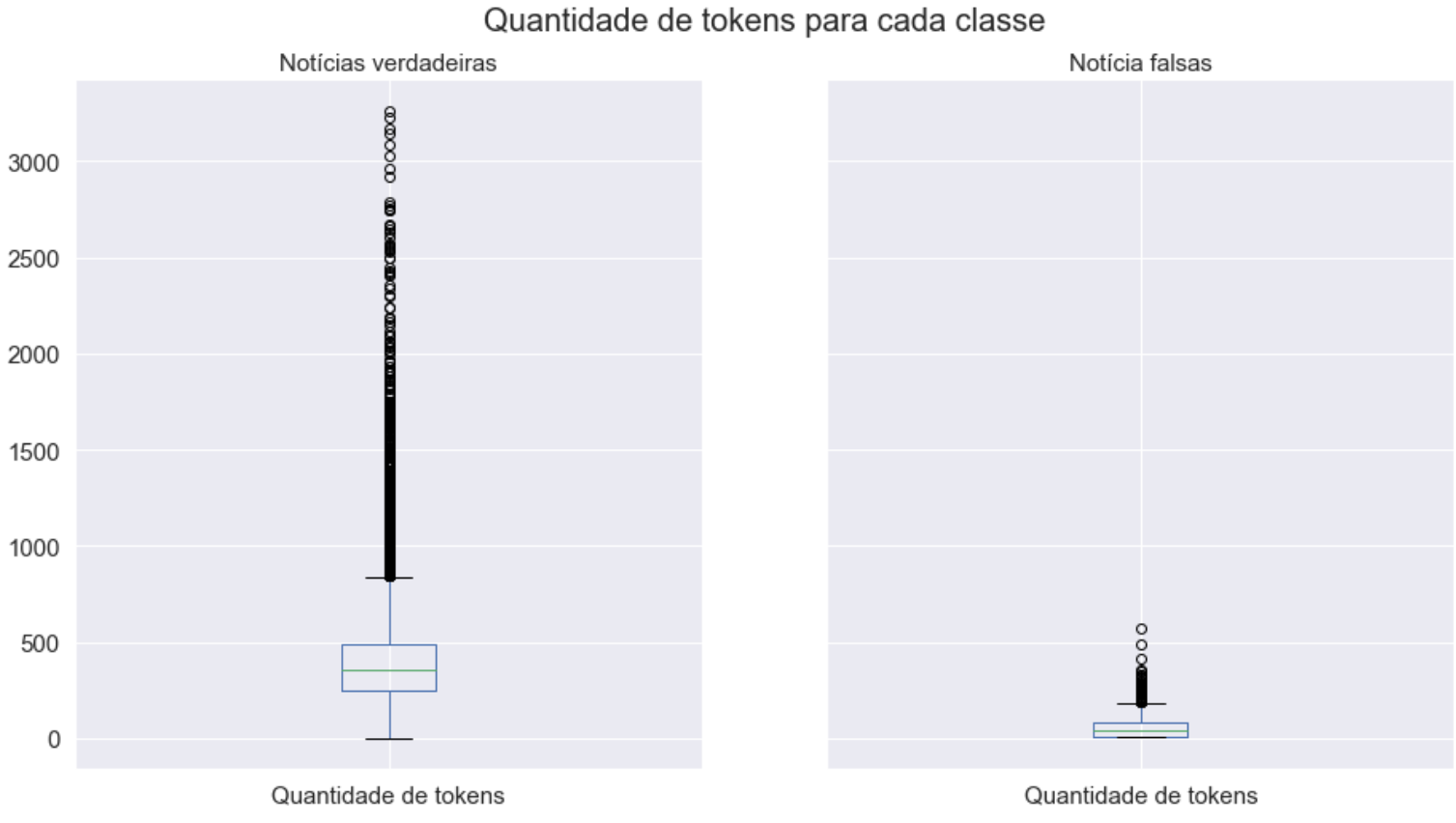
- Aumento de 1153 para 1251 registros (98 novos)

ANÁLISE DOS DADOS

ANÁLISE

- Quantidade de tokens por classe de documento;
- Tokens mais comuns para cada classe;
- Wordclouds;
- Redução de dimensionalidade (PCA & t-SNE);
- LIME.

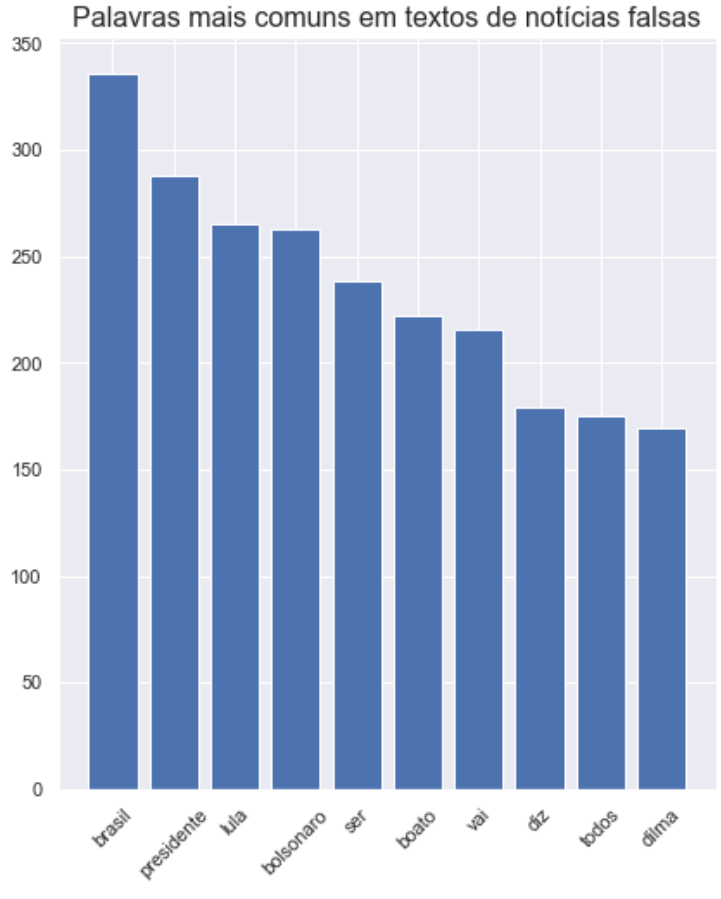
BOXPLOT

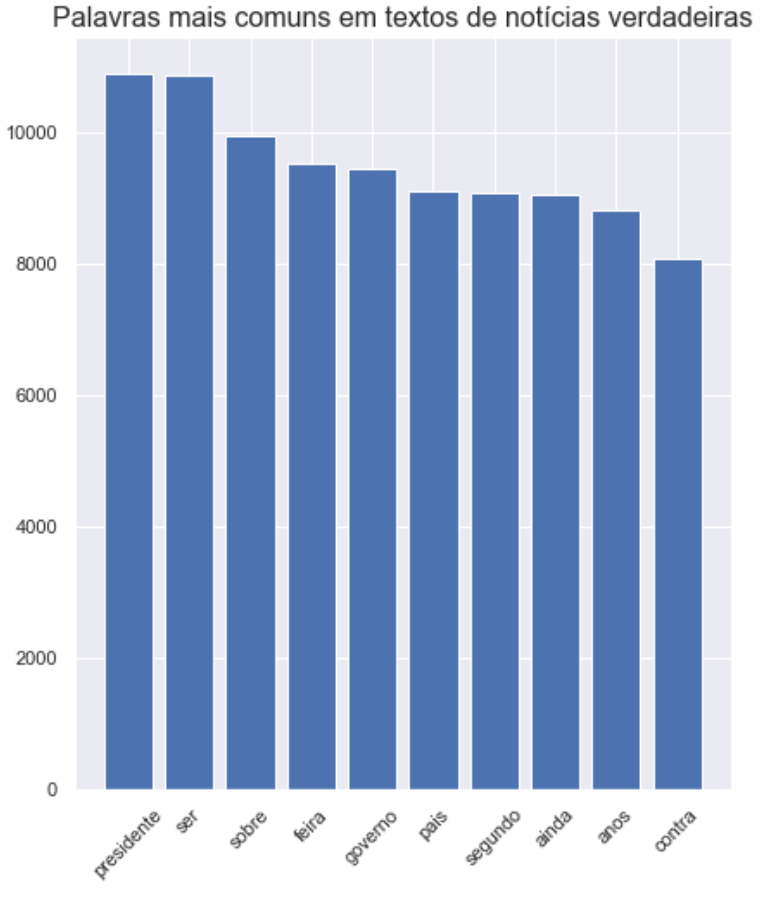

Quantidade de tokens para cada classe
Notícias verdadeiras
Notícia falsas
3000
2500
2000
1500
1000
500
0
Quantidade de tokens
Quantidade de tokens

WORDCLOUD - NOTÍCIAS FALSAS

WORDCLOUD - NOTÍCIAS VERDADEIRAS

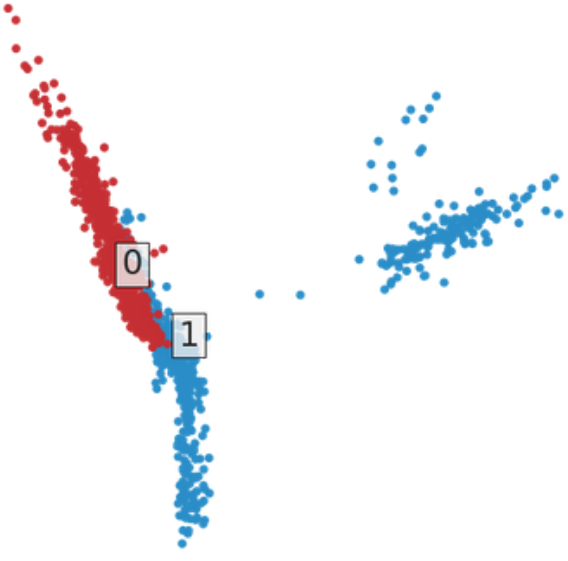
BARPLOT


Palavras mais comuns em textos de notícias falsas
350
300
250
200
150
100
50
0
brasil
presidente
lula
bolsonaro
ser
boato
vai
diz
todos
dilma

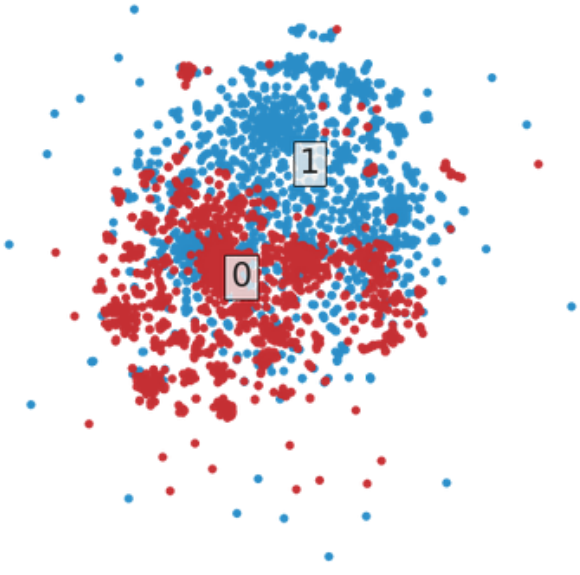

Palavras mais comuns em textos de notícias verdadeiras
10000
8000
6000
4000
2000
0
presidente
ser
sobre
feira
governo
país
segundo
ainda
anos
contra

REDUÇÃO DE DIMENSIONALIDADE

PCA

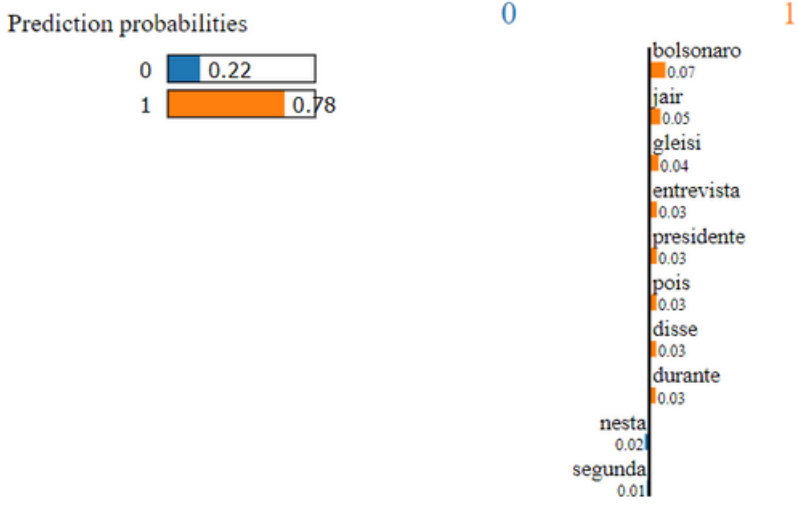


t-SNE



LIME

Tenho 78% de certeza de que a notícia é falsa.



Text with highlighted words

entra acao stf impedir bolsonaro reduza preco gas acordo fonte dentro congresso integrantes partido trabalhadores enviaram nesta segunda feira stf acao exigindo presidente jair bolsonaro impedido prosseguir projeto reducao preco gas pais medida presidente inconstitucional objetivo unico aumentar popularidade perante populacao pobre bolsonaro parceria ministro economia paulo guedes pretende reduzir preco gas domestico meses medidas ministro melhorar economia pais curto medio prazo presidente gleisi hoffmann afirmou partido fara tudo preco gas continue acabamos enviar acao stf barre medida populista descabida desse fascista reducao preco gas trara grandes prejuizos querida petrobras cofres

PIPELINE DE NLP

LIMPEZA DOS TEXTOS

- Importância da limpeza:
  - Sujeira;
  - Tokenização.
- O que é removido:
  - Acentos;
  - Tokens que intercalam letras e números;
  - Caracteres não-alfabéticos;
  - Tokens com letras que aparecem mais de 2 vezes em seguida;
  - Palavras de tamanho menor ou igual a 2;
  - Stopwords.

VETORIZAÇÃO

- Representação do texto em números;
- **TF-IDF:**
  - A importância da palavra **aumenta** de acordo com a ocorrência dela no documento, mas é **penalizada** pela ocorrência dela no corpus.
  - **TF** (Term Frequency): quantas vezes um token aparece num determinado documento;
  - **IDF** (Inverse Document Frequency): calculado a partir de quantos documentos no corpus têm o token;
  - Atingiu uma **acurácia** de ~97%.

VETORIZAÇÃO

- Word2Vec:
  - Rede neural;
  - Gera um espaço vetorial:
    - Palavras em similares ficam próximas.
  - Ex.: Rei - Homem + Mulher = Rainha
  - Atingiu uma **acurácia** de ~92%

```
Palavra: lula
Palavras semelhantes:
- petista
- dilma
- lula_preso
- haddad
- pupila
- partido_trabalhadores
- inacio_lula
- dilma_rousseff
- lula_silva
- juiz_moro
```

```
Palavra: bolsonaro
Palavras semelhantes:
- jair_bolsonaro
- bolsonaro_psl
- mourao
- presidente_eleito
- capitao_reformado
- capitao_reserva
- eduardo_bolsonaro
- candidato_psl
- bolsonarista
- bolsonaro_psc
```

MODELO

Logistic Regression	0.975
Naive Bayes	0.645
Random Forest	**0.978**
Support Vector Machine	0.727
XGBoost	0.962
Neural Networks	0.977

ARQUITETURA DA APLICAÇÃO

ARQUITETURA

ionic

SANIC FRAMEWORK

OBRIGADO!