IDENTIFICAÇÃO DE AUTORIA EM TEXTOS CURTOS UTILIZANDO TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL

João Otávio Rodrigues Ferreira Frediani
Orientador: Prof. Assoc. Aparecido Nilceu Marana

Introdução

Processamento de linguagem natural

Área que estuda a utilização de computadores para compreender e manipular linguagens humanas.

A área busca solucionar tarefas como a de tradução, reconhecimento de fala, sumarização, análise de sentimentos, análise de autoria entre outros

Análise de autoria

Análise de autoria é uma sub-área do processamento de linguagem natural que engloba as tarefas:

- Busca de características sociolinguísticas do autor
- Autenticação de autoria
- Atribuição de autoria

Atribuição de autoria

Atribuição de autoria é a tarefa que busca identificar o autor de um texto dentro de um grupo de autores conhecidos.

- Utilizada para diferentes propósitos
- Historicamente aplicada em textos longos

Textos na internet atual

- Textos na internet são usualmente curtos

- Twitter limita em 280 caracteres por publicação

Problema atualmente

Em 2015 o New York Times revelou a existência de uma empresa russa especializada na propagação de informações falsas.

Em 2017 o Reino Unido ameaçou sanções a redes sociais como Facebook e Twitter que não auxiliassem na identificação da origem de notícias falsas.

Em 2021 um ex-funcionário anônimo do Facebook revelou que a rede social ignorava publicações que continham discurso de ódio pois estas geravam engajamento.

Objetivo

O objetivo deste trabalho é avaliar diferentes modelos de processamento de linguagem natural na tarefa de atribuição de autoria para textos retirados da rede social Twitter.

O trabalho procurou implementar modelos que utilizam de técnicas clássicas de processamento de linguagem natural e um modelo de aprendizado de máquina para observar as diferenças em seu desempenho.

Extração de característica

Bag of Words

- Dados precisam ser convertidos para formatos compreensíveis ao computador.

- É uma técnica clássica de representação de palavras.

- Cada frase é armazenada como um dicionário não ordenado.

Frase 1: Ontem comi uma torta de maçã

Frase 2: Eu comi uma bela maçã

Dicionário

0	1	2	3	4	5	6	7
Ontem	comi	uma	torta	de	maçã	eu	bela

Frase 3: Comi uma torta de frango deliciosa

Representação Bag of Words

[0 , 1, 1, 1, 1, 0, 0, 0]

N-gram

- Consiste em separar o texto em fatias de tamanhos iguais.

- N-grams podem ser feitos para palavras ou caracteres.

Nível de caractere:

Frase : “Bom diaaa !!!”

2-gram : {“Bo”, “om”, “m ”, “ d”, “di” “ia”, “aa”, “aa”, “a ”, “ !”, “!!”, “!!”}

Nível de palavra:

Frase : “Estava escuro então acendi uma vela”

3-gram : {“Estava escuro então”, “escuro então acendi”, “então acendi uma”, “acendi uma vela”}

TF-IDF

- Medida de frequência aplicada à uma sequência específica
- O valor é dado por uma relação que diminui a medida que uma palavra é utilizada pelo corpus e aumenta quanto mais usada em um texto.

$$p_d = f_{p,d} * log(\frac{|D|}{f_{p,D}})$$

Word Embeddings

- Busca evitar generalizações comuns em outros métodos.

- Palavras são representadas por vetores em que cada posição relaciona um valor numérico à uma palavra

Ex:

$$REI = [0.32 \quad 0.12 \quad 0.98 \quad 0.56 \quad 0.68]$$

Rei - Homem = Realeza

Realeza + Mulher = Rainha

Classificadores

Naive Bayes

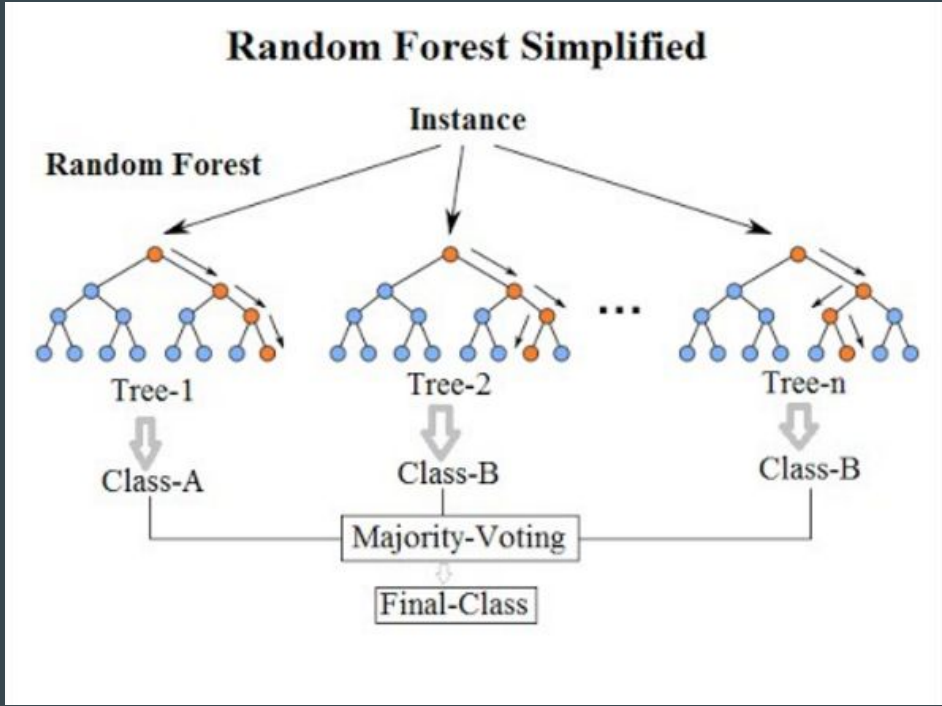
- Algoritmo probabilístico de aprendizado supervisionado baseado na aplicação do teorema de Bayes.
- O classificador é chamado de naive (ingênuo) pois supõe que os vetores de características são independentes.

$$P(c|X) = \frac{P(X|c)P(c)}{P(X)}$$

$$F_b(X) = arg \max_{c \in C} P(X|c)P(c)$$

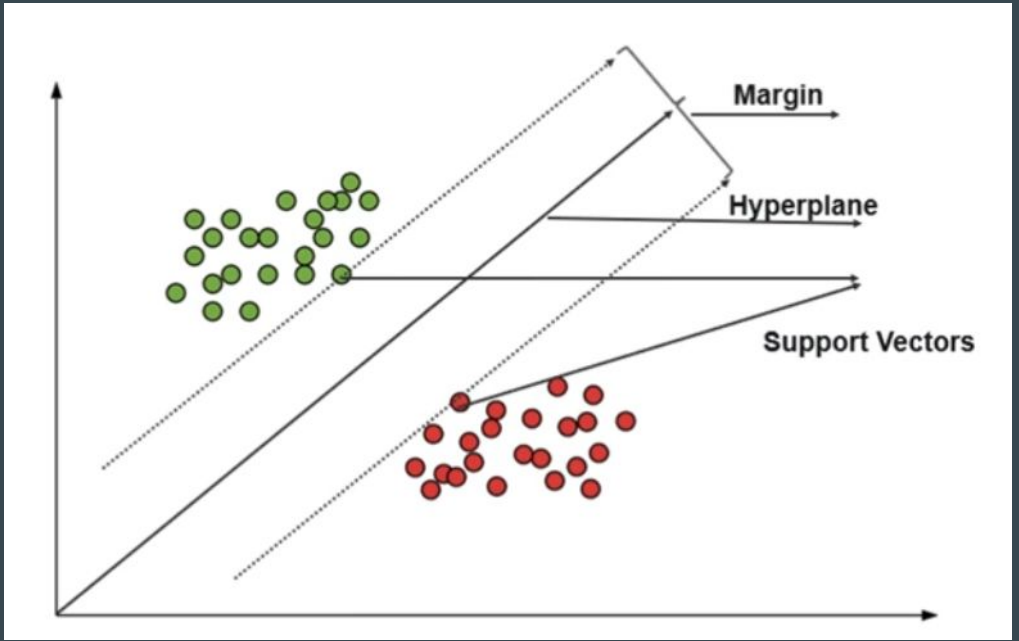
Random Forest

- O classificador de floresta aleatória consiste simplificadamente da união de árvores de decisão.
- Uma árvore de decisão busca valores que separam uma classe da outra para alguma característica.



Support Vector Machine

- Máquinas de vetores de suporte são classificadores que buscam separar os dados em classes procurando um hiperplano que descreva a fronteira entre elas.
- Problemas de multiclasse são tratados como diversos problemas binários

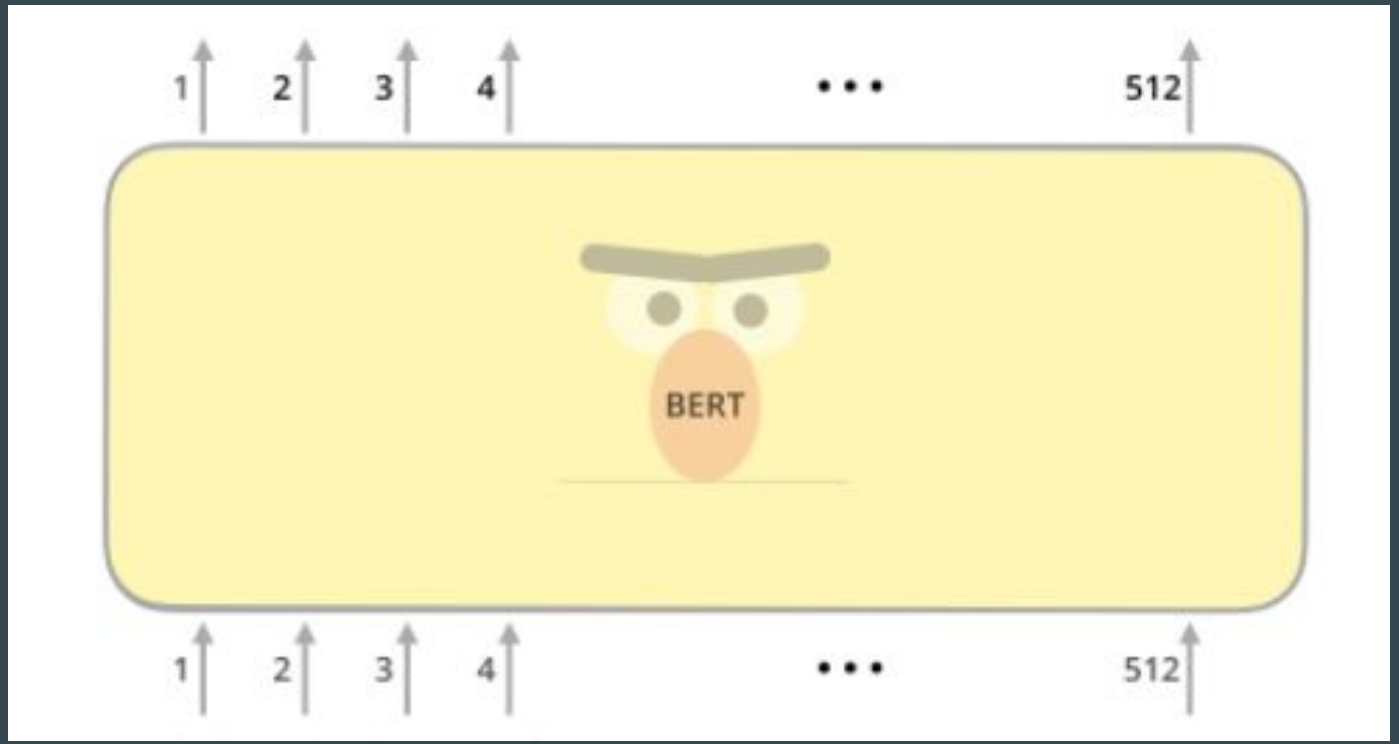


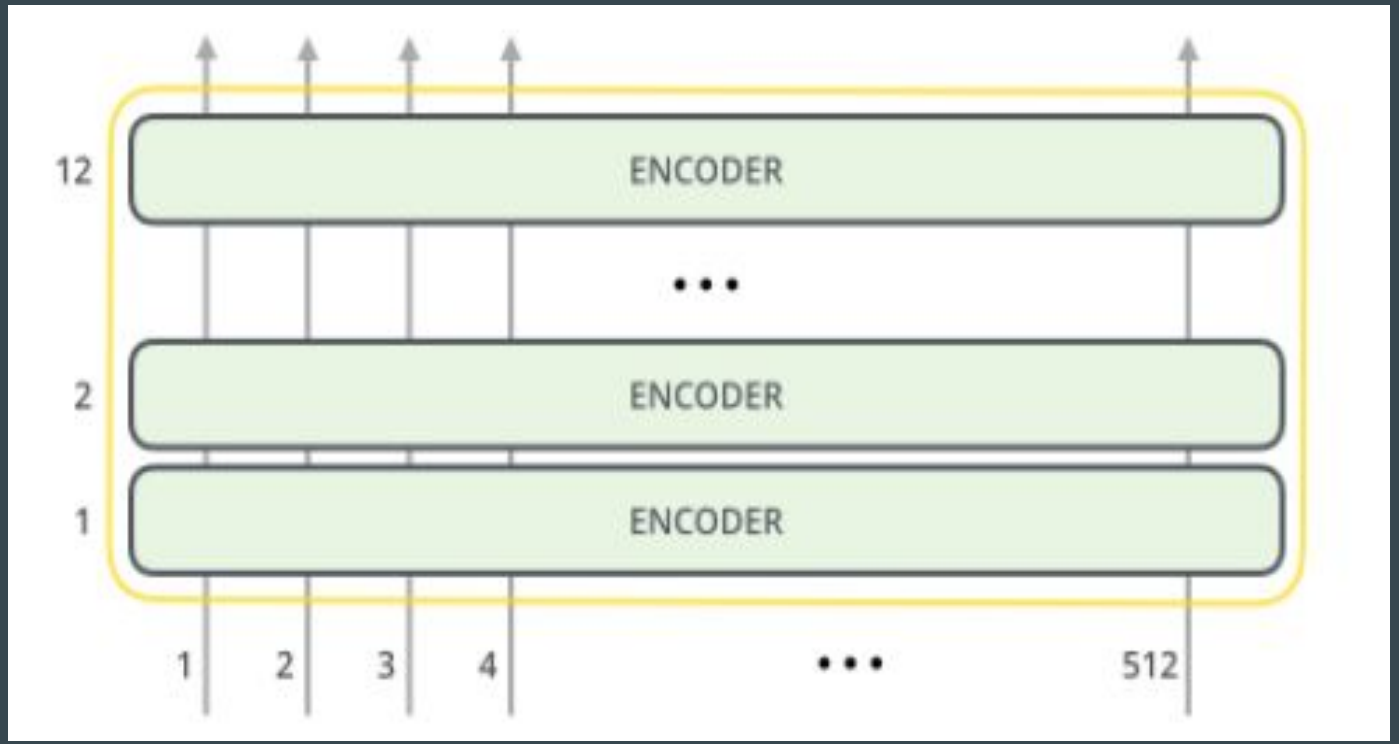
Modelo BERT

BERT

- Modelo proposto por pesquisadores da equipe de inteligência artificial para textos da Google em 2018.

- O modelo é baseado em técnicas desenvolvidas em diversos outros trabalhos como Transformers e aprendizado semi-supervisionado


1
2
3
4
• • •
512
BERT
1
2
3
4
• • •
512


12
ENCODER
2
ENCODER
1
ENCODER
1
2
3
4
512

Pré-treinamento

O modelo BERT é pré-treinado em duas tarefas baseadas em máscaras:

"Deus ajuda quem cedo madruga"

Para compreender palavras por contexto bilateral:

"Deus ajuda quem [máscara] madruga"

Para compreender relação entre frases:

"Deus ajuda quem"+"cedo madruga"

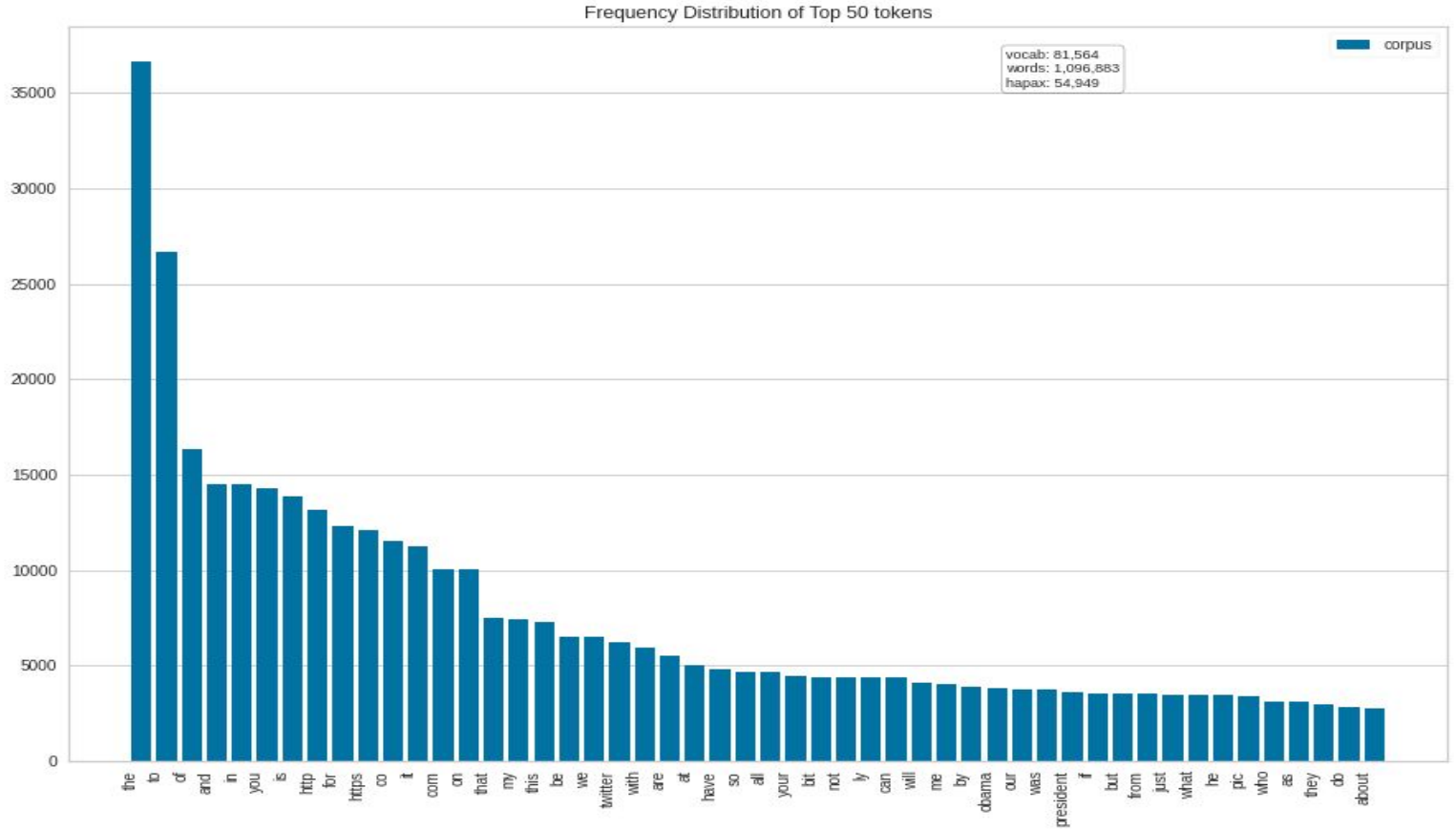
"Deus ajuda quem"+"não se olha os dentes"

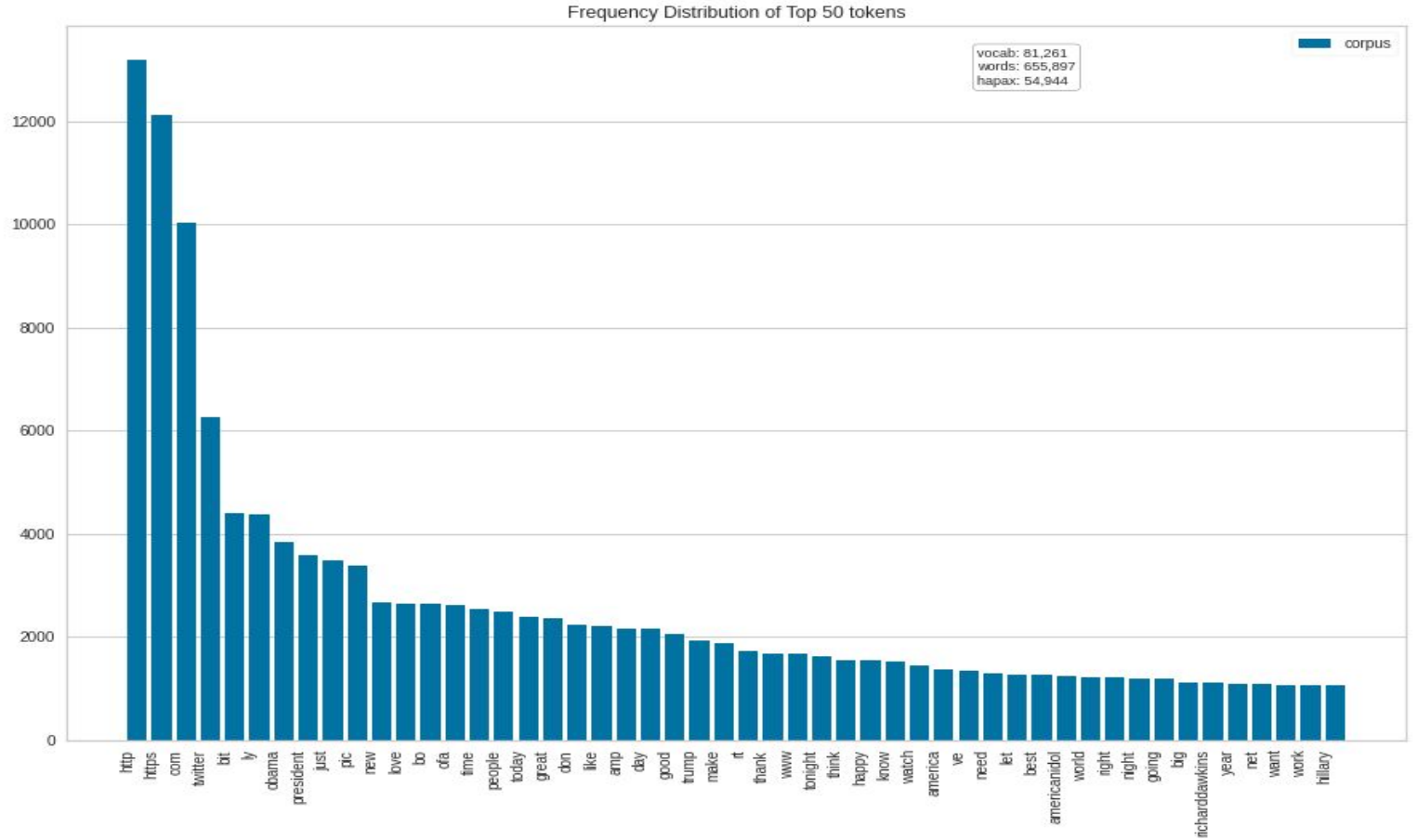
Dados

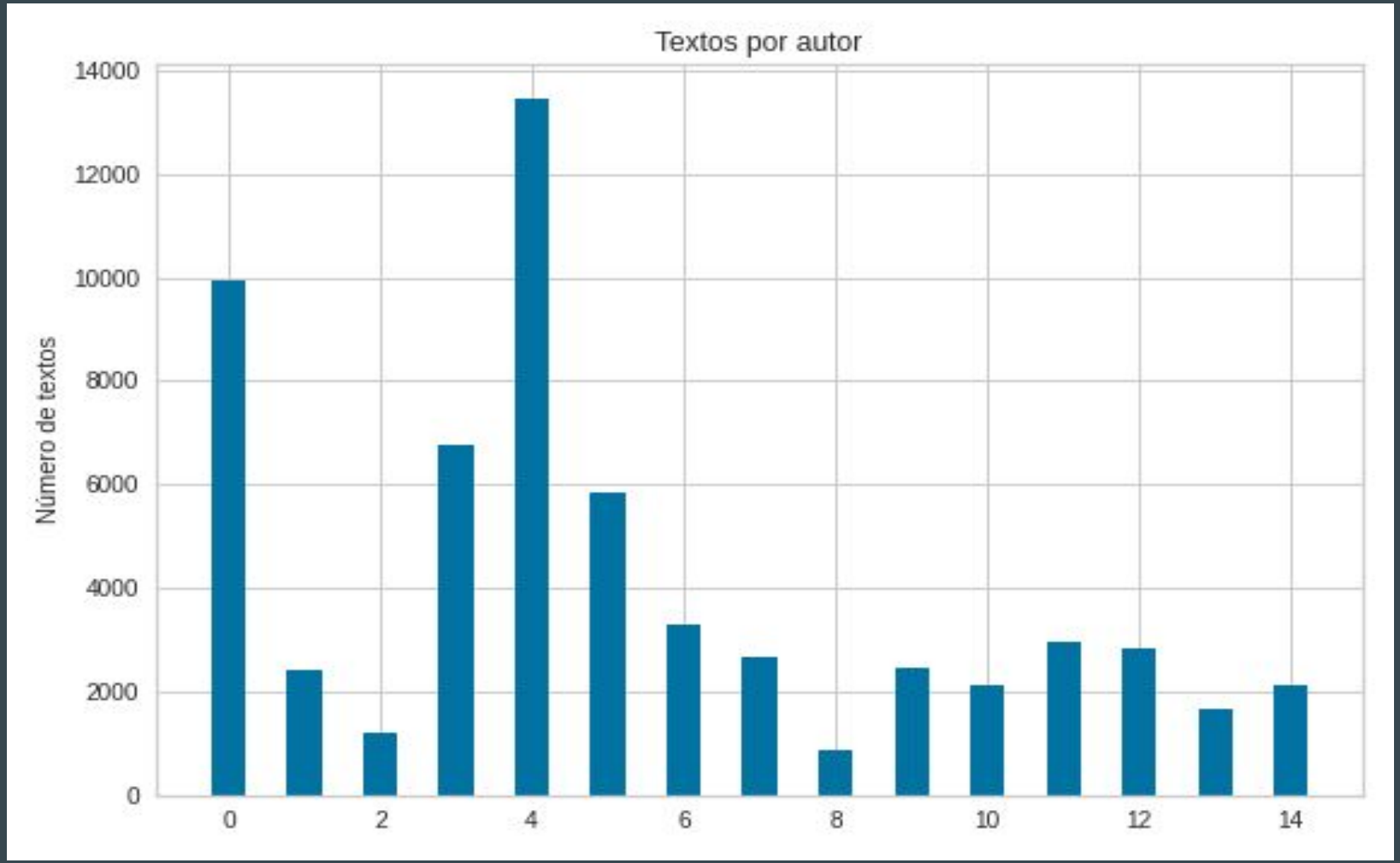
Corpus

- Resultado da união de dados encontrados no site *Kaggle* com dados obtidos utilizando a *Twitter API*.

- Os dados são então *Tweets* de 15 diferentes autores escritos em inglês, 7 destes autores foram recolhidos do *Kaggle* e os outros 8 foram colhidos pela API.


Frequency Distribution of Top 50 tokens
corpus
vocab: 81,564
words: 1,096,883
hapax: 54,949
35000
30000
25000
20000
15000
10000
5000
0
the
to
of
and
in
you
is
http
for
https
co
it
com
on
that
my
this
be
we
twitter
with
are
at
have
so
all
your
bit
not
ly
can
will
me
by
obama
our
was
president
if
but
from
just
what
he
pic
who
as
they
do
about


Frequency Distribution of Top 50 tokens
corpus
vocab: 81,261
words: 655,897
hapax: 54,944
12000
10000
8000
6000
4000
2000
0
http
https
com
twitter
bit
ly
obama
president
just
pic
new
love
bo
dfa
time
people
today
great
don
like
amp
day
good
trump
make
rt
thank
www
tonight
think
happy
know
watch
america
ve
need
let
best
americanidol
world
right
night
going
big
richarddawkins
year
net
want
work
hillary


Textos por autor
Número de textos
14000
12000
10000
8000
6000
4000
2000
0
0
2
4
6
8
10
12
14

Pré-Processamento

- Links são substituídos por “URL”
- Menções a outros usuários são substituídas por “REF”
- Datas são substituídas por “DATE”
- Horários são substituídos por “TIME”
- Números são substituídos por “NUM

Desenvolvimento

Ferramentas

Modelos

Os modelos implementados são resultado da combinação das técnicas de extração de características mais o modelo BERT

Técnicas de extração de características:

- TF-IDF + Unigram (n-gram a nível de palavra de tamanho 1).
- TF-IDF + 4-gram a nível de caractere.
- TF-IDF + 4-gram a nível de palavra.
- TF-IDF + 4-gram a nível de caractere + n-gram a nível de palavra de tamanho 1 á 5

Classificadores:

- Random Forest.
- Naive Bayes.
- Support Vector Machine.

Treinamento

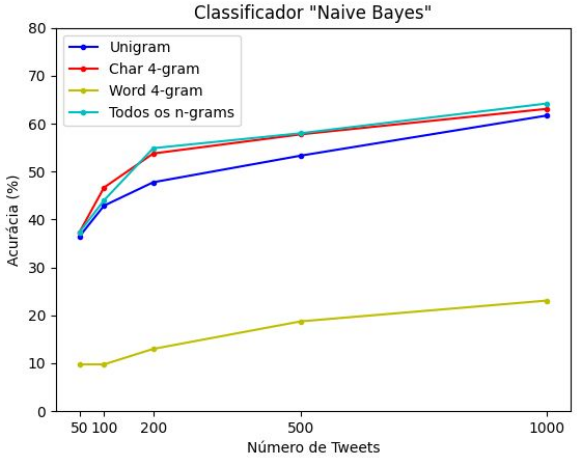
- Os modelos foram treinados para 50, 100, 200, 500 e 1000 textos por autor.
- Para os modelos clássicos foram utilizadas implementações disponibilizadas pela biblioteca Scikit learn e para o modelo BERT foi utilizada a versão pré-treinada para categorização de sequência disponibilizada pela biblioteca Hugging Face realizando o processo de fine-tuning.
- Os dados foram separados em uma proporção 70/30, 70% para treinamento e 30% para teste.

Teste

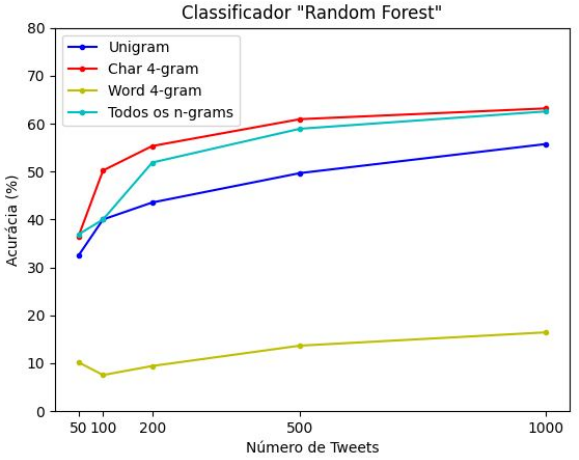
- Os modelos foram então testados com 30% dos dados.

- O modelo que obteve os melhores resultados foi o modelo BERT treinado com 700 tweets por autor e testado com 300 com acurácia de 75.78% e pontuação F1 0.7590.

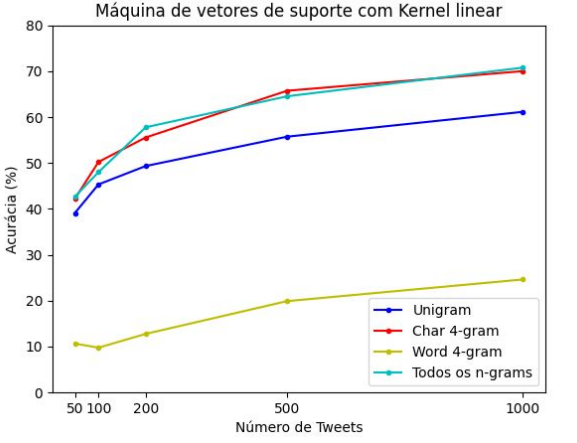
- Dentre somente os modelos clássicos o melhor desempenho foi o do modelo com todas as técnicas de extração de características com classificador SVM que obteve acurácia de 70.77% e pontuação F1 0.7099



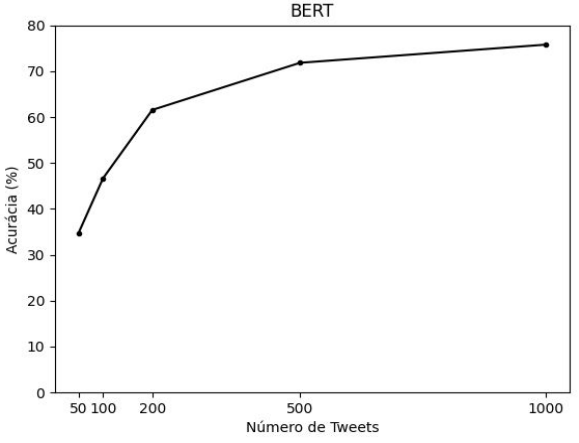
Melhor desempenho: 64.21%



Melhor desempenho: 63.18%

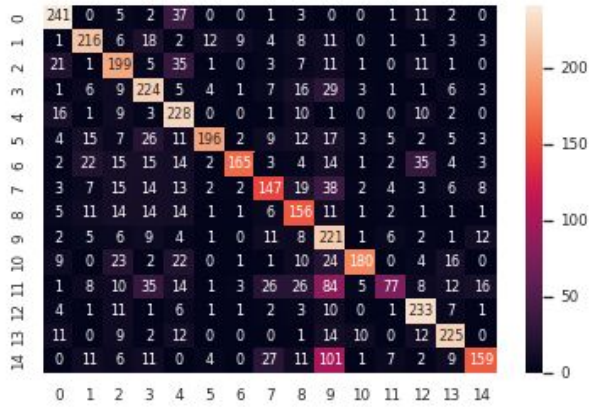


Melhor desempenho: 70.77%

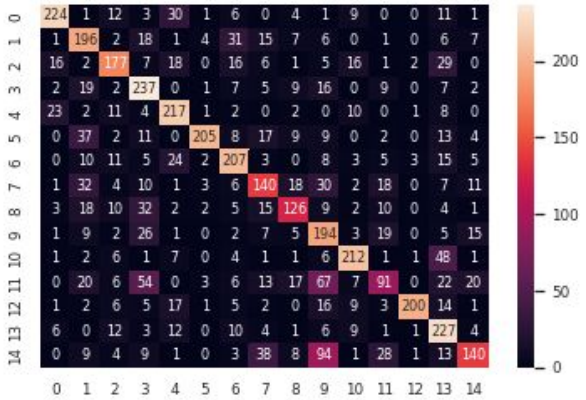


Melhor desempenho: 75.78%

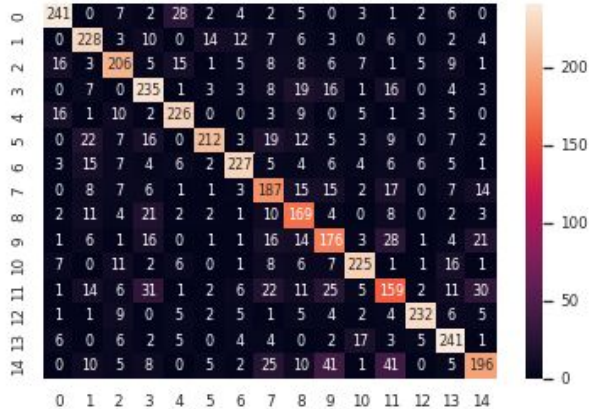
Modelo NB



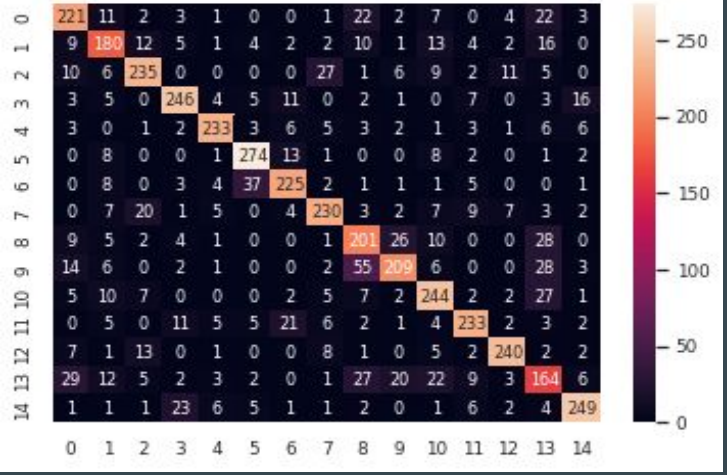
Modelo RF

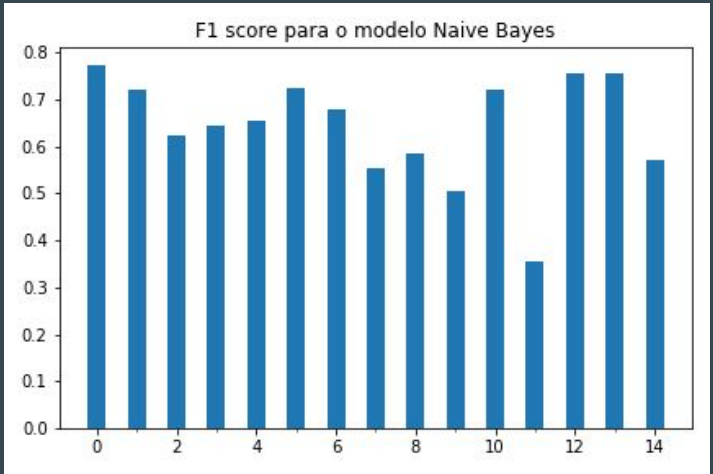


Modelo SVM

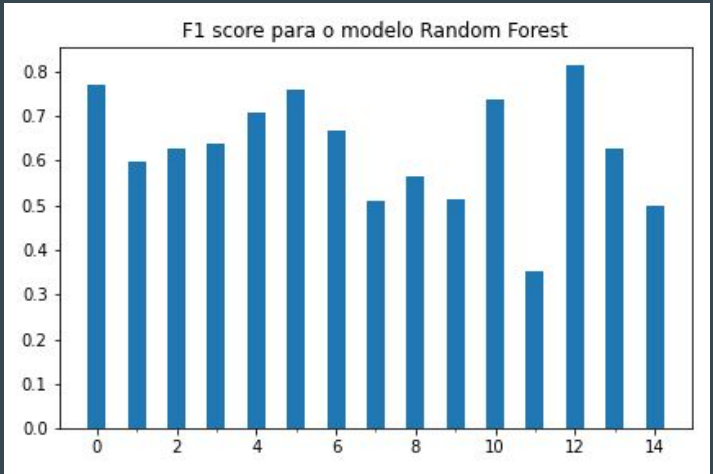


Modelo BERT

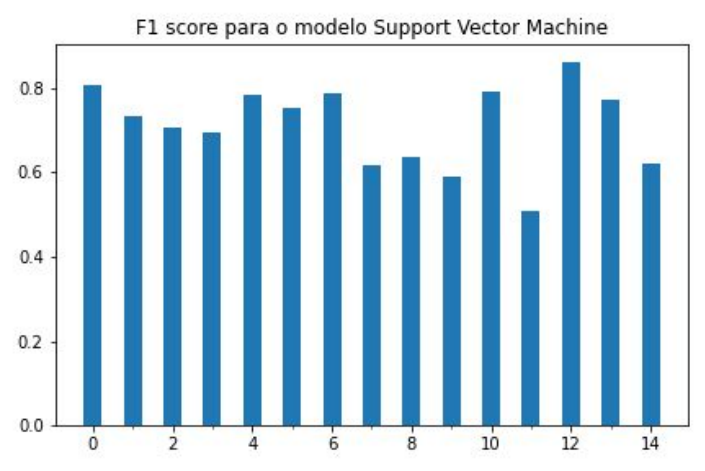




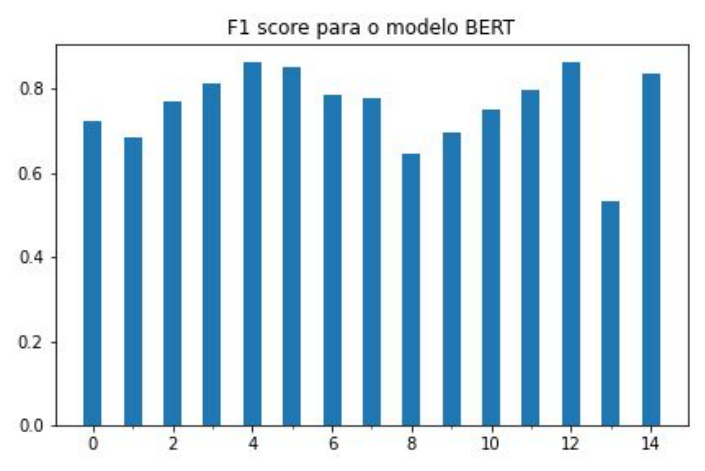
F1 = 0.6408



F1 = 0.6261



F1 = 0.7099



F1 = 0.7590

Conclusão

- A utilização de inteligências artificiais pode ser uma ferramenta auxiliar na busca de autores de textos criminosos na internet.
- Textos retirados da internet apresentam diversos desafios únicos.
- Comparação no desempenho de modelos clássicos e um modelo no estado da arte.

Bibliografia

ABORISADE, O.; ANWAR, M. Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers. In: IEEE. 2018 IEEE International Conference on Information Reuse and Integration (IRI). [S.l.], 2018. p. 269–276.

CHOWDHURY, G. Natural language processing. Annual Review of Information Science and Technology, v. 37, n. 1, p. 51–89, jan. 2003. ISSN 0066-4200.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

ROCHA, A.; SCHEIRER, W. J.; FORSTALL, C. W.; CAVALCANTE, T.; THEOPHILO, A.; SHEN, B.; CARVALHO, A. R.; STAMATATOS, E. Authorship attribution for social media forensics. IEEE transactions on information forensics and security, IEEE, v. 12, n. 1, p. 5–33, 2016.

WOLF, T.; DEBUT, L.; SANH, V.; CHAUMOND, J.; DELANGUE, C.; MOI, A.; CISTAC, P.; RAULT, T.; LOUF, R.; FUNTOWICZ, M. et al. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771, 2019.

CHENG, A. The Agency. 2015. Https://www.nytimes.com/2015/06/07/magazine/theagency.html Acessado em 09/02/2022.

HERN, A. Facebook and Twitter threatened with sanctions in UK ’fake news’ inquiry. 2017. Https://www.theguardian.com/media/2017/dec/28/facebook-and-twitter-threatened-withsanctions-in-uk-fake-news-inquiry Acessado em 09/02/2022.

TIMBERG, C. New whistleblower claims Facebook allowed hate, illegal activity to go unchecked. 2021. Https://www.washingtonpost.com/technology/2021/10/22/facebook-newwhistleblower-complaint/ Acessado em 09/02/2022.

Muito Obrigado!