A Web of Hate: Tackling Hateful Speech in Online Social Spaces

Processamento de Linguagem Natural

Integrantes

Caique de Camargo — **11091312**

caique.camargo@aluno.ufabc.edu.br

Jean Augusto — **21025614**

jean.a@aluno.ufabc.edu.br

July Anne Pinheiro — **11094013**

july.pinheiro@aluno.ufabc.edu.br

Marcela Yamashita — **21083913**

marcela.a@aluno.ufabc.edu.br

Introdução

Artigo: A Web of Hate: Tackling Hateful Speech in Online Social Spaces

Autores: H. M. Saleem, K. P. Dillon, S. Benesch, D. Ruths

46 citações. Apresentado no First Workshop on Text Analytics for Cybersecurity and Online Safety at LREC (2016)

Proposta do artigo:
- Comparar abordagens diferentes para definição de discurso de ódio
- Propor uma nova abordagem para detecção de discurso de ódio

Community based x Keyword based

Community based x Keyword based

Community based: utilização da linguagem que emerge de grupos que se auto-organizam (como Reddit). Grupos se formam através de práticas linguísticas comuns. **“the group is defined by speech and the speech comes to define the group”**

Keyword based: classificação baseada em palavras-chave

Proposta do grupo: keyword-based

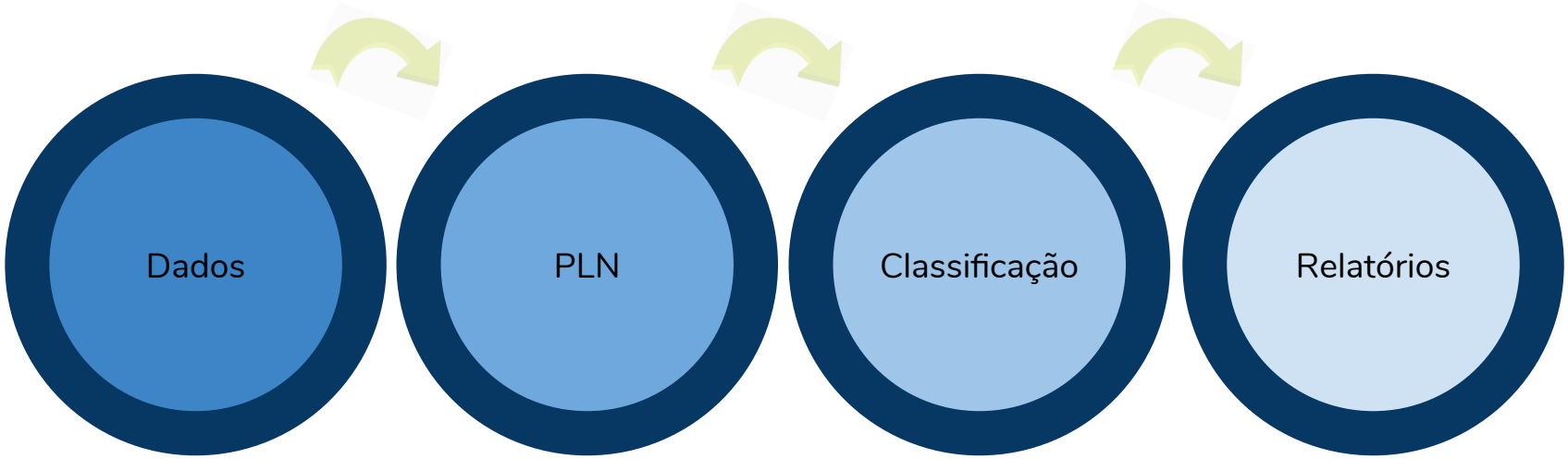
Inabilidade de encontrar datasets de grupos fechados

Testar a performance de diferentes classificadores (Naive Bayes, SVM e LR) no dataset do Twitter

Twitter: ~8 milhões de usuários no Brasil (julho/2019)

Grande influência no cenário social e político.

Organização



Base de dados

Original

- Stuck_In_The_Matrix Reddit Dataset (250GB compress.)
- Voat
- Web forums

Adaptação

- “Hate Speech Identification” (Twitter Dataset) - Data World (cerca de 15mil tweets)

Algoritmos

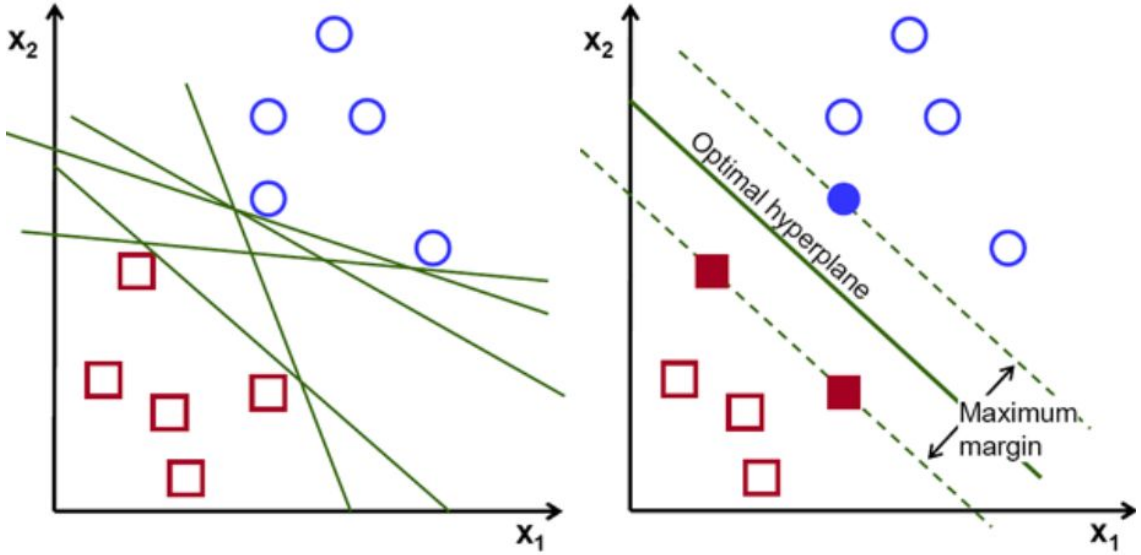
Implementação e comparação de resultados de três algoritmos de classificação:

- Naive Bayes
- Support Vector Machine (SVM)
- Logistic Regression (LR)

Naive Bayes

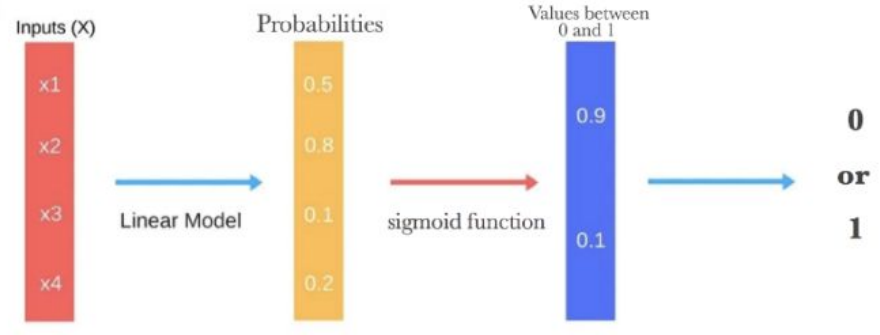
$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

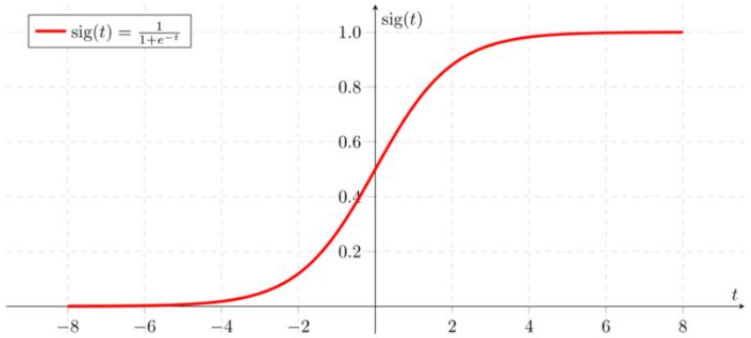
SVM



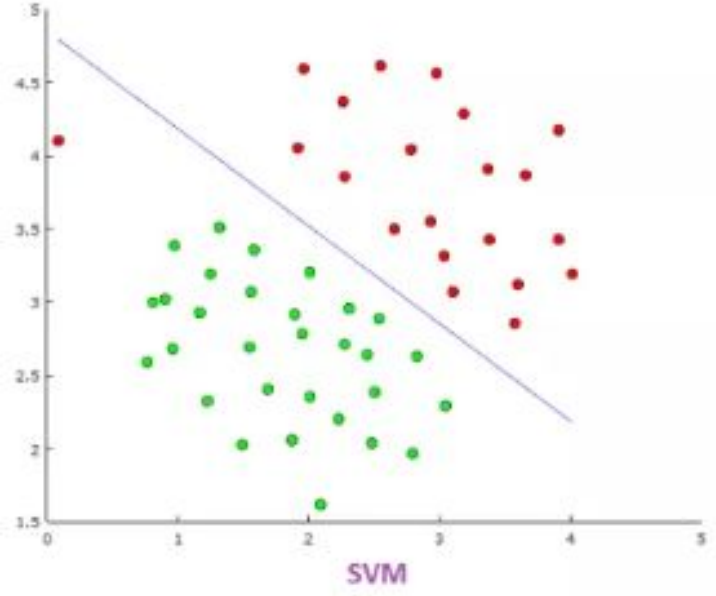
Obs: Utilização do kernel linear do SVM.

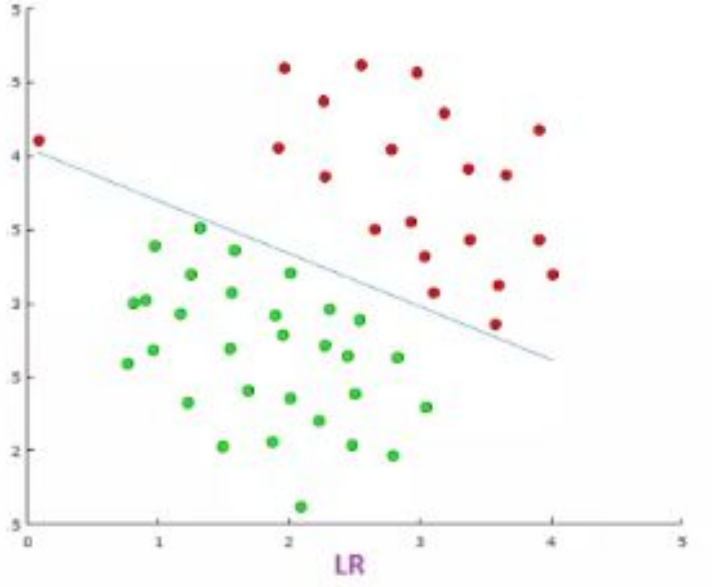
Logistic Regression



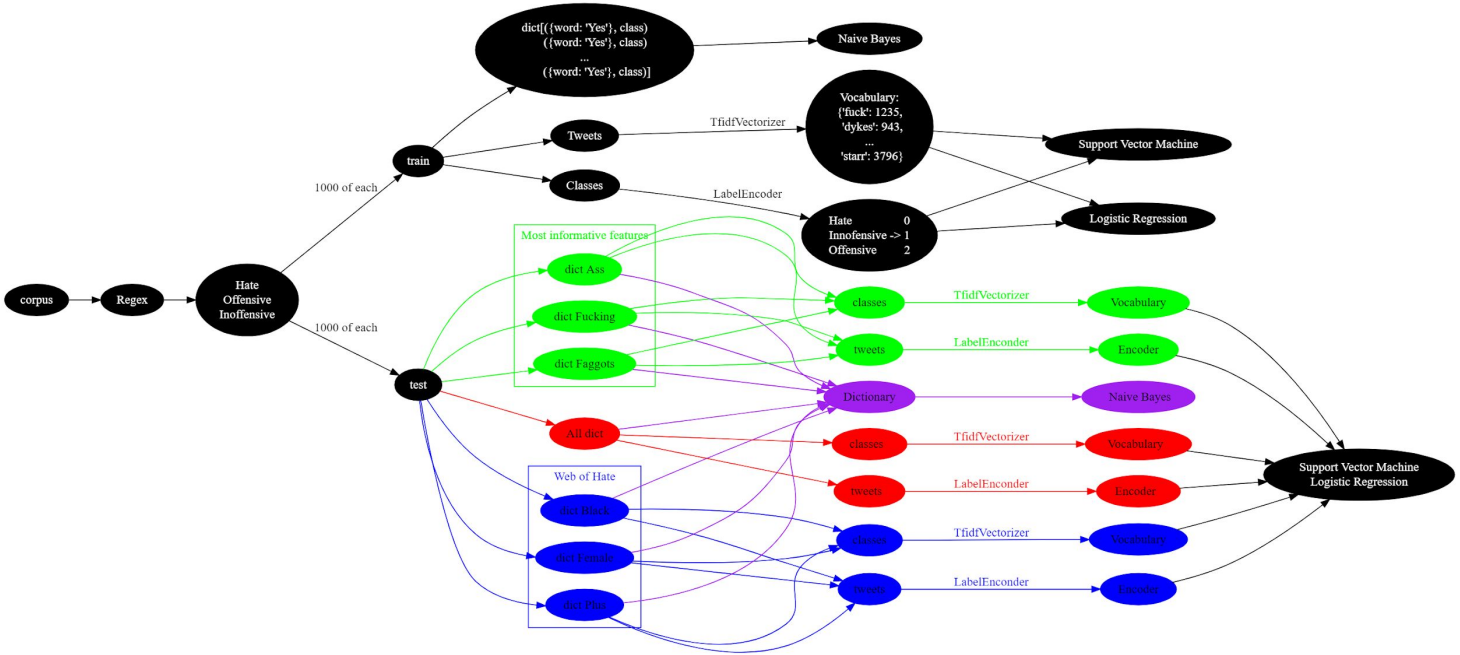


SVM X LR

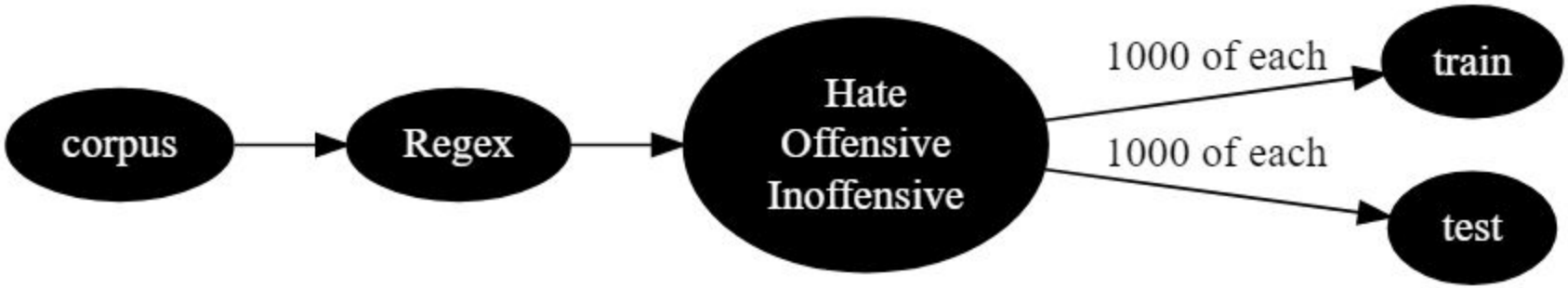




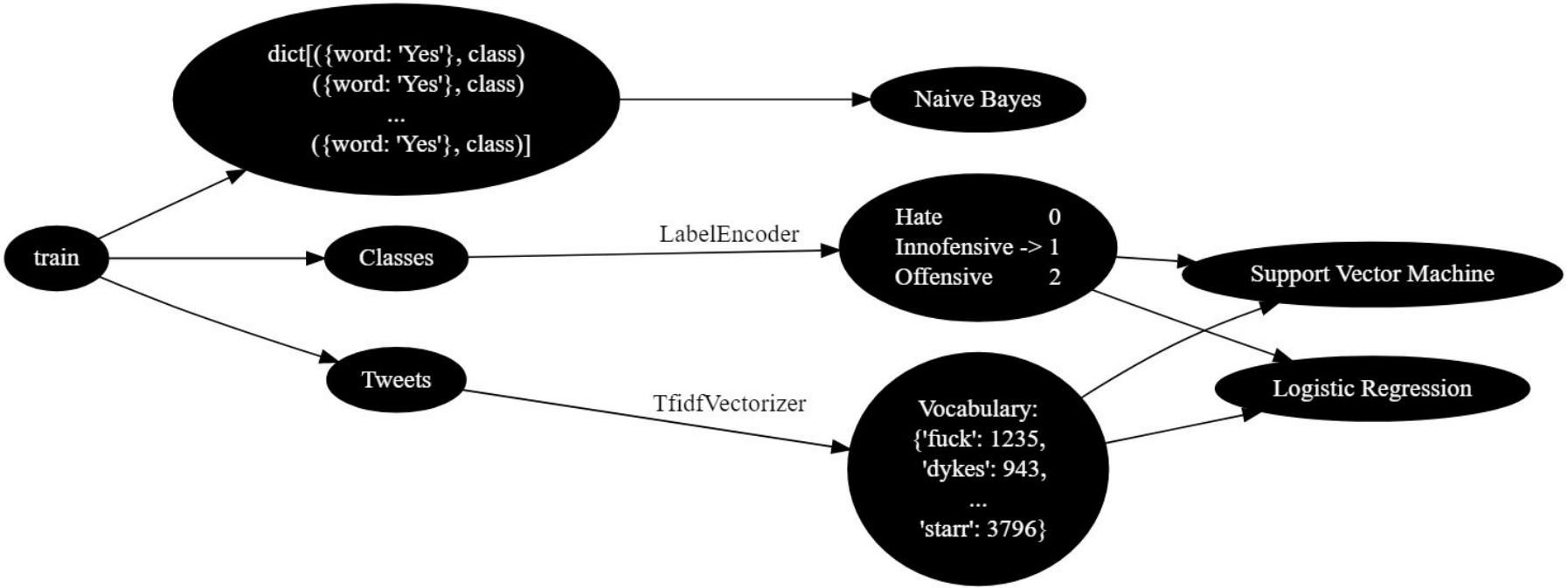
Algoritmo


corpus
Regex
Hate
Offensive
Inoffensive
1000 of each
1000 of each
train
test
dict[{word: 'Yes'}, class)
({word: 'Yes'}, class)
...
({word: 'Yes'}, class)]
Naive Bayes
Tweets
Classes
TfidfVectorizer
LabelEncoder
Vocabulary:
{'fuck': 1235,
'dykes': 943,
...
'starr': 3796}
Hate 0
Innofensive -> 1
Offensive 2
Support Vector Machine
Logistic Regression
Most informative features
dict Ass
dict Fucking
dict Faggots
All dict
Web of Hate
dict Black
dict Female
dict Piss
classes
tweets
Dictionary
classes
tweets
classes
tweets
TfidfVectorizer
LabelEnconder
TfidfVectorizer
LabelEnconder
TfidfVectorizer
LabelEnconder
Vocabulary
Encoder
Naive Bayes
Vocabulary
Encoder
Vocabulary
Encoder
Support Vector Machine
Logistic Regression

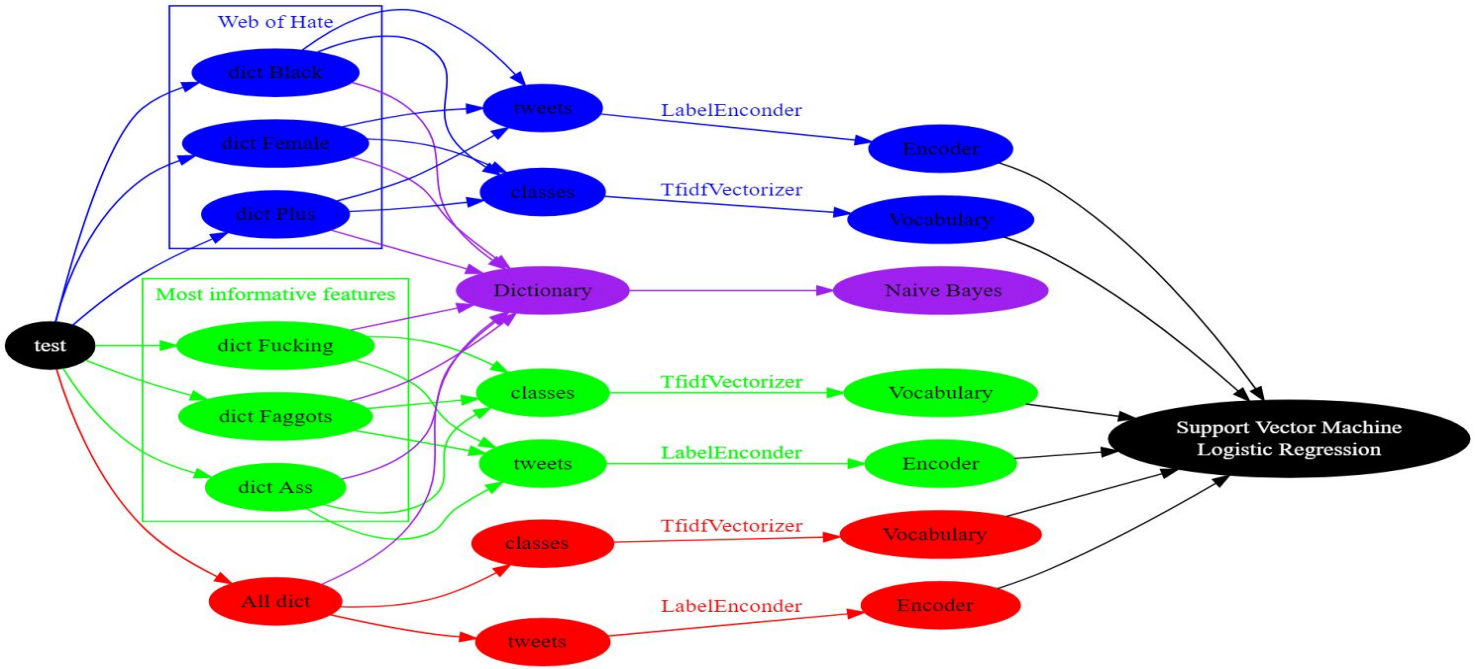
Algoritmo


corpus
Regex
Hate
Offensive
Inoffensive
1000 of each
train
1000 of each
test

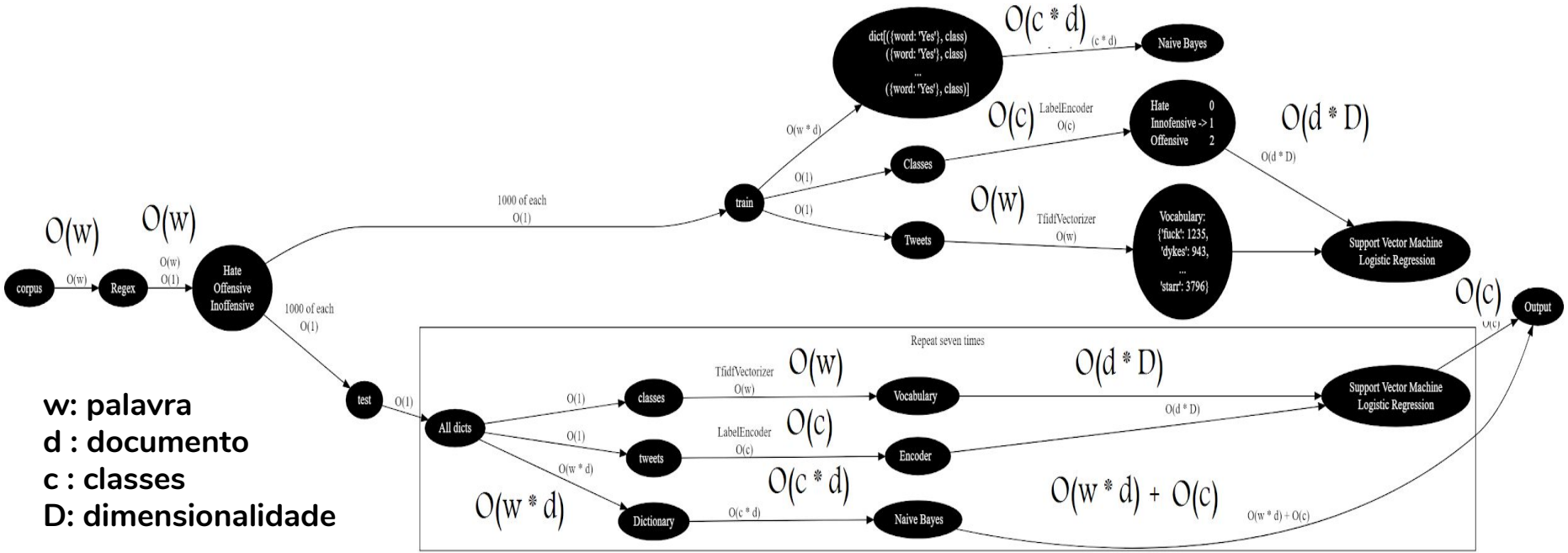
Algoritmo


dict[({word: 'Yes'}, class)
({word: 'Yes'}, class)
...
({word: 'Yes'}, class)]
Naive Bayes
train
Classes
LabelEncoder
Hate 0
Innofensive -> 1
Offensive 2
Support Vector Machine
Tweets
TfidfVectorizer
Vocabulary:
{'fuck': 1235,
'dykes': 943,
...
'starr': 3796}
Logistic Regression

Algoritmo


Web of Hate
dict Black
dict Female
dict Plus
Most informative features
dict Fucking
dict Faggots
dict Ass
test
All dict
tweets
classes
Dictionary
classes
tweets
classes
tweets
LabelEnconder
TfidfVectorizer
Naive Bayes
TfidfVectorizer
LabelEnconder
TfidfVectorizer
LabelEnconder
Encoder
Vocabulary
Vocabulary
Encoder
Vocabulary
Encoder
Support Vector Machine
Logistic Regression

Complexidade



w: palavra
d : documento
c : classes
D: dimensionalidade

O(11w + 14c + 7d + 14(w*d) + 8(c*d) + 9(d*D))

Resultados

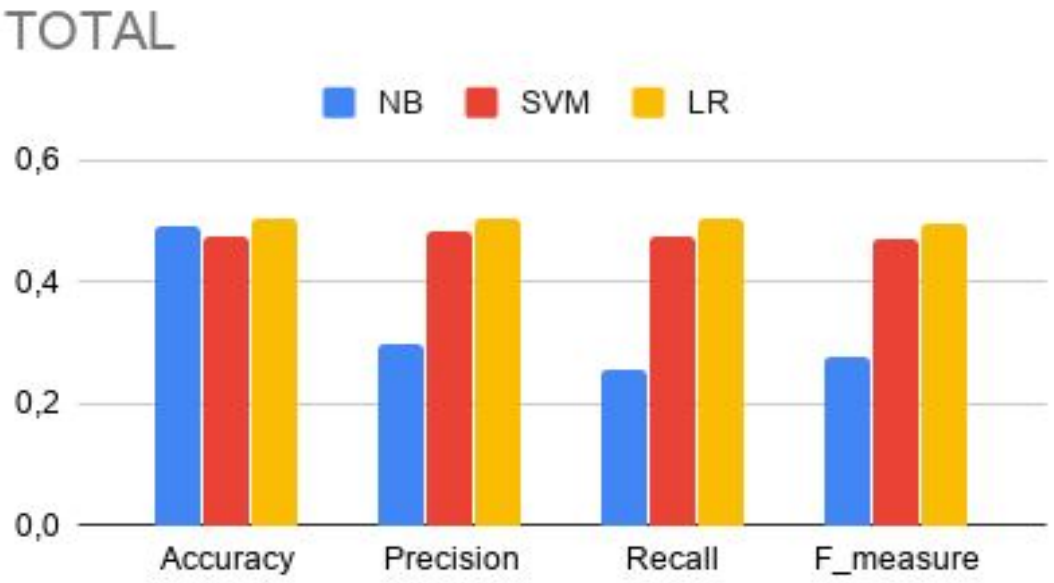
Medidas:

- Accuracy
- Precision
- Recall
- F_measure

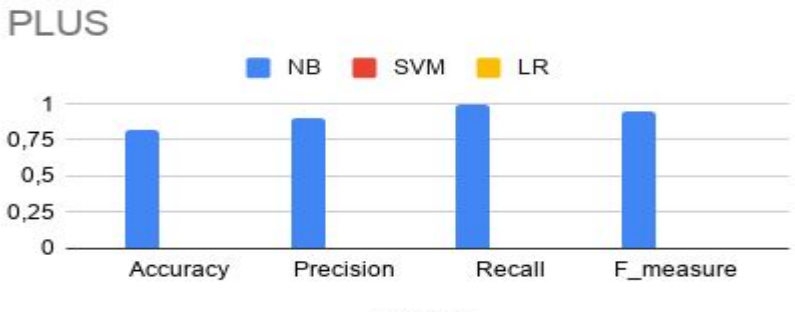
Termos utilizados:

- Black
- Plus
- Female

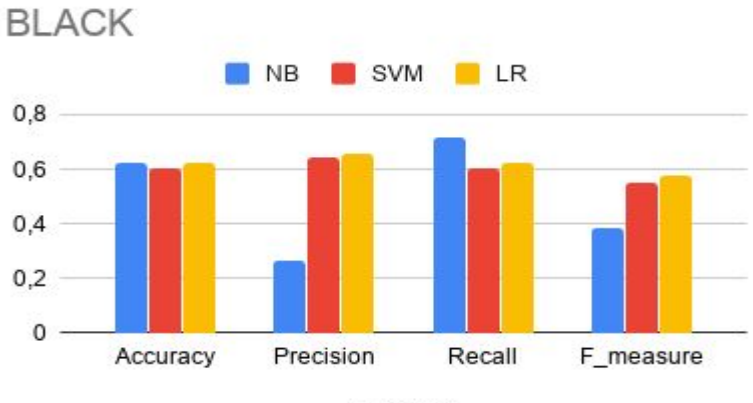
Resultados


TOTAL
NB
SVM
LR
0,6
0,4
0,2
0,0
Accuracy
Precision
Recall
F_measure

Resultados

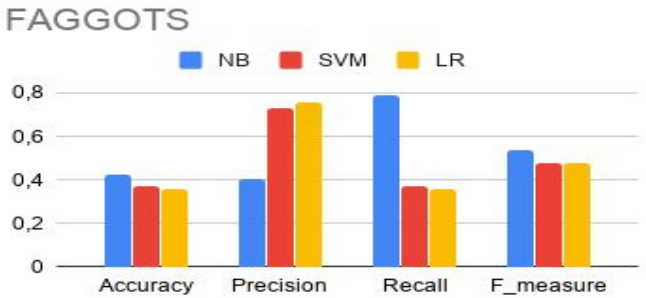

PLUS
NB
SVM
LR
1
0,75
0,5
0,25
0
Accuracy
Precision
Recall
F_measure


FEMALE
NB
SVM
LR
0,6
0,4
0,2
0
Accuracy
Precision
Recall
F_measure


BLACK
NB
SVM
LR
0,8
0,6
0,4
0,2
0
Accuracy
Precision
Recall
F_measure

Resultados


FUCKING
NB
SVM
LR
0,8
0,6
0,4
0,2
0
Accuracy
Precision
Recall
F_measure


FAGGOTS
NB
SVM
LR
0,8
0,6
0,4
0,2
0
Accuracy
Precision
Recall
F_measure


ASS
NB
SVM
LR
0,6
0,4
0,2
0,0
Accuracy
Precision
Recall
F_measure

Obrigado!