



# Naive Bayes

**João Seike 9784634**

**Rafael Nazima 11208311**

**Daniel Adan Pereira 11295866**



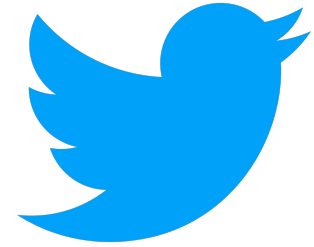
A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines, rendered in a light gray color.

1.

# Pré-processamento dos dados

# Conjunto de dados

- © 1.578.627 tweets.
- © Rótulo 1 para sentimento positivo 😊
- © Rótulo 0 para sentimento negativo 😞



# Construção do vocabulário

- © Estrutura auxiliar contendo todas as palavras (Linked List).
- © Limpeza dos dados (remoção de termos neutros):
  - Busca Binária ❌
  - HashSet ✅
- © Guardamos a frequência das palavras em tweets positivos e em tweets negativos, assim sabemos o quanto cada palavra influencia positivamente ou negativamente. Essa lista é guardada como serializable em um arquivo **.ser**

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by small circles, some of which are larger and have concentric circles, suggesting a hierarchical or multi-layered structure. The lines are thin and gray, connecting the nodes in a non-linear fashion.

# 2. **Experimentos**

# Experimentos

Embaralhamos toda vez o SentimentAnalysisData



Realizamos o Holdout, CrossValidation, e Holdout( usando stop word).

Utilizando do mesmo SentimentAnalysisData embaralhado podemos comparar facilmente os resultados.

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by small circles, some of which are larger and have concentric circles, suggesting a hierarchical or multi-layered structure. The lines are thin and gray, connecting the nodes in a non-linear fashion.

# 3. Resultados

# Cross-validation

**Erro Padrao 10-fold :**

0.0003546877024790768

**Erro Verdadeiro vai ficar entre :**

0.27256191514073885 e 0.2739522909344569

Fold	Erro	Acertos	Acurácia
1	43343	114512	0.72542523
2	42856	114999	0.72851034
3	43343	114512	0.72542523
4	43021	114834	0.72746508
5	43231	114624	0.72613474
6	43177	114678	0.72647683
7	43390	114465	0.72512749
8	42999	114856	0.72760445
9	42912	114943	0.72815559
10	43043	114812	0.72732571



# Holdout

## Dados originais:

- © Acertos: 380053
- © Erros: 146131

## Dados com *stop words* removidas

- © Acertos: 376058
- © Erros: 150127



# Holdout

Matriz Confusão:

		Avaliado	
		Sentimento Positivo	Sentimento Negativo
Tweet	Sentimento Positivo	154520	109000
	Sentimento Negativo	36507	226156

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by small circles, some of which are larger and have concentric rings, suggesting a hierarchical or multi-layered structure. The lines are thin and gray, connecting the nodes in a non-linear fashion.

# 4. Conclusões



“

- ◎ Taxa de erros consistente para todos os testes.
- ◎ Grande velocidade de processamento.
- ◎ Simplicidade do algoritmo

# Obrigado!