CIÊNCIA DE DADOS (BIG DATA PROCESSING AND ANALYTICS)

RECUPERAÇÃO DA INFORMAÇÃO NA WEB E EM REDES SOCIAIS

Professor curador
Luciano Moreira Camilo e Silva

TRILHA 6
PROCESSAMENTO DE LINGUAGEM NATURAL
ANÁLISE DE SENTIMENTOS

O que são sentimentos?

"Fiquei até as 3h estudando! Valeu a pena, amo aprender coisas novas."

- São **individuais**, **subjetivos** e **privados**.
- Ainda assim, podem ser **verdadeiros**.
- Expresso de forma **explicita** ou **implícita**.

  - *"Meu novo celular é incrível!"*
  - "O fone durou apenas dois meses."

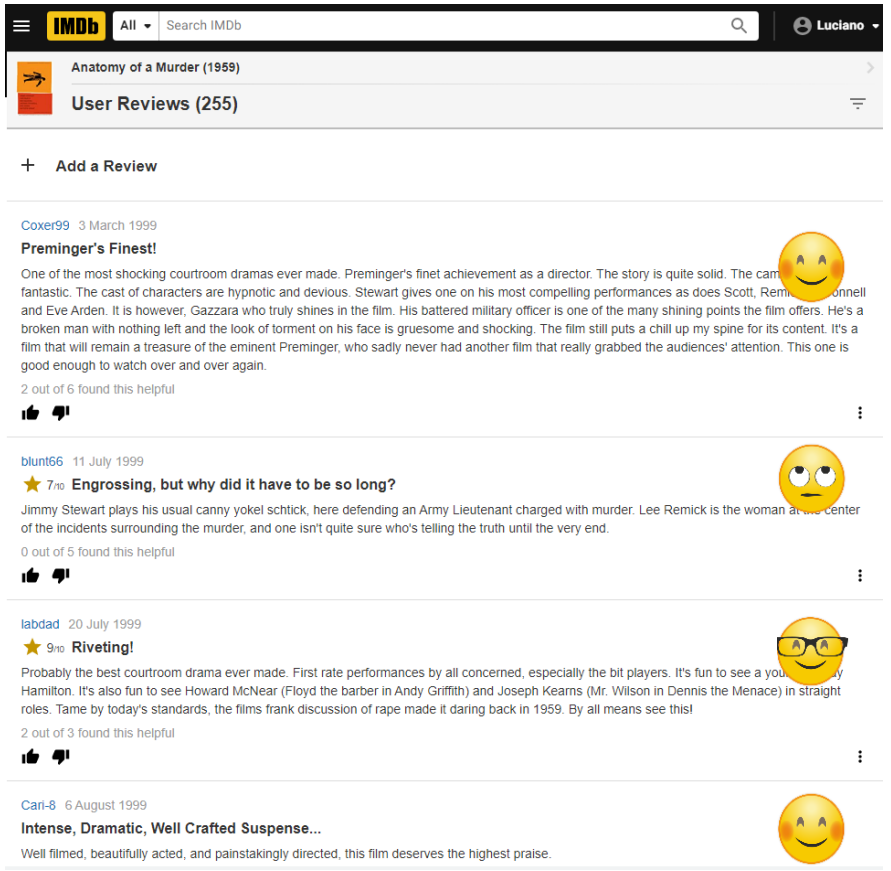
Fonte: Imagem de Pixaline por Pixabay

O que são análises de sentimentos?

- Subárea de **ciência da computação**, dentro da disciplina de **Processamento de Linguagem Natural**.

- Também conhecida como **mineração de textos** ou **extração de avaliações**.

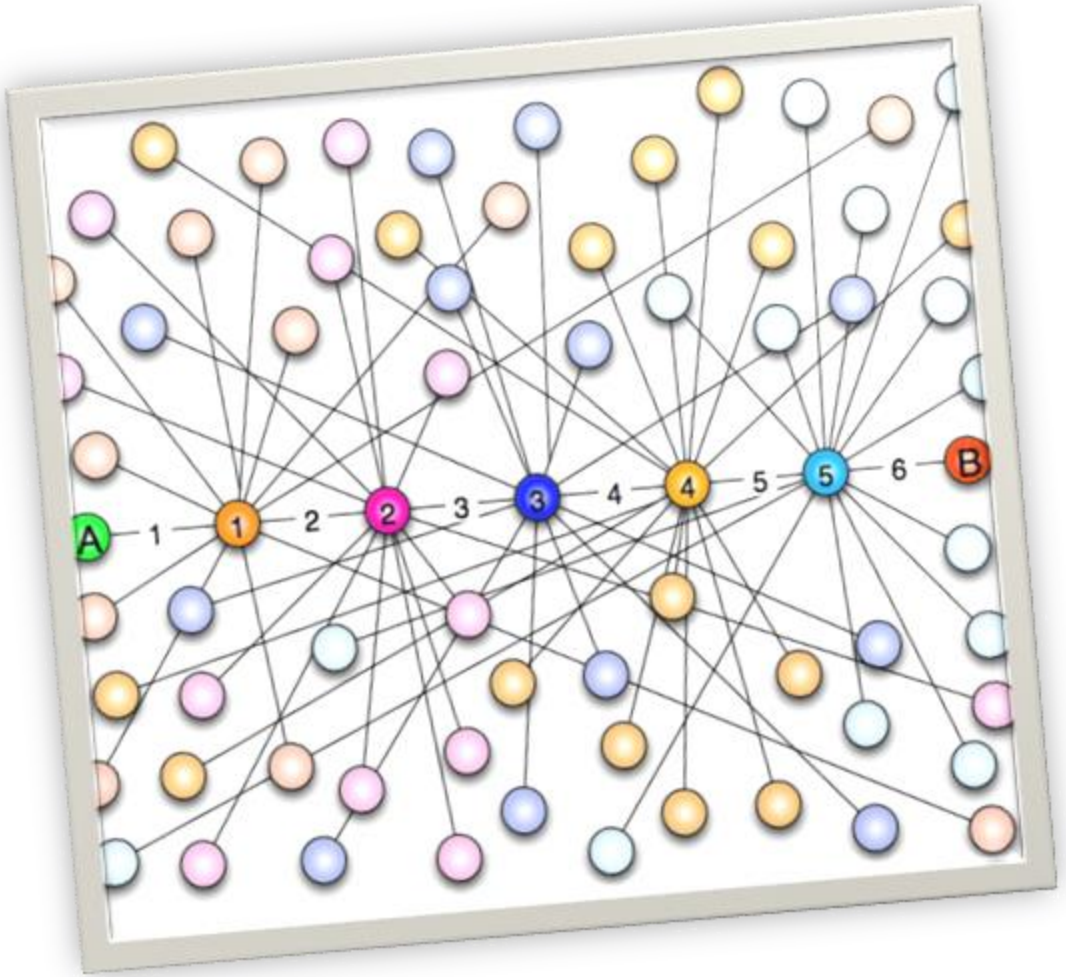
- Utilizada em monitoramento de marca, eleições e políticas públicas etc.



Fonte: Internet Movie Database

O que são (sistemas de) rede social?

- As pessoas se identificam a partir de um **perfil**, real ou virtual.
- Conecta pessoas **simétrica** (Facebook) e **assimetricamente** (Twitter).
- Expansão das relações humanas por meio de **conexões virtuais**
  (# relações virtuais > # relações presenciais).
- **Conteúdo gerado** ou compartilhado pelos **próprios participantes** da rede.



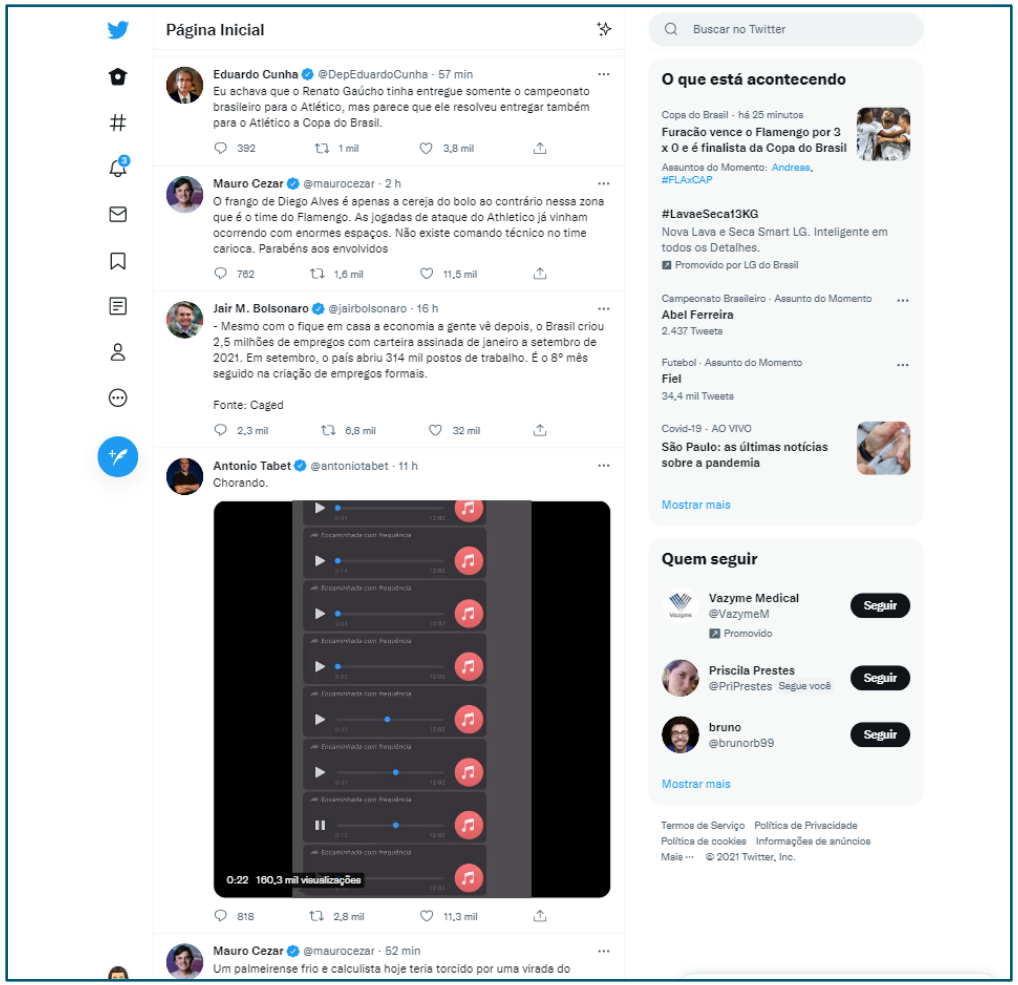
Linha do tempo das redes s

1997 1999 2002 2003 2004 2005 2006 2007 2009 2010 2011 2012 2013 2015 2016 2020

CYWORLD

e: Elaborada pelo autor.

Twitter



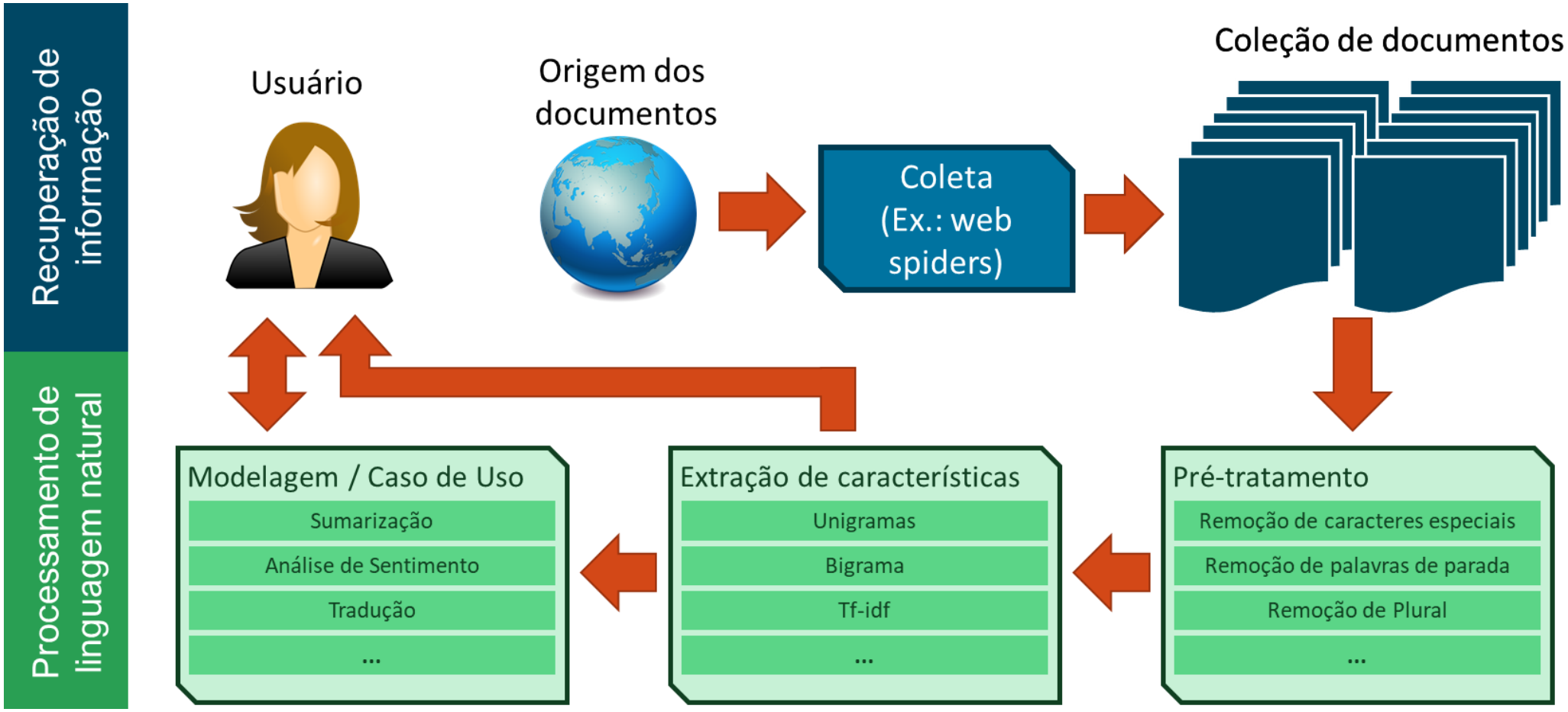
- Fundado em 2006.
- +200 MM de usuários ativos e bilhões de mensagens trocadas mensalmente.
- Popular para comunicação rápida e expressão de opinião.

Amostragem dos tweets coletados

	id_str	text	screen_name	search_term	search_term_group	source
3583	1450267147377201152	Puta merda chegou Maligno na hbo max mas quem ...	taisouzax	hbo max	hbomax	Twitter for iPhone
4191	1452387178592411648	Os lançamentos da Netflix em novembro de 2021 ...	canaltech	netflix	netflix	canalte.ch
8009	1449871350915637248	Uma obra prima \nhttps://t.co/hU16a4q1Oy	muhkent	primevideo	primevideo	Twitter for Android
5056	1451326165713670144	Sacrifício é ter que aguentar o Luciano Hulk n...	hatakerenata	netflix	netflix	Twitter for Android
5757	1450876023554330624	@alleywaysx @SemideusGrecia @danielbfx @marcos...	sr_albertini	netflix	netflix	Twitter for Android
1051	1450589824709185536	@dudadopke Com dólar tem Disney, com bitcoin t...	brurichards	disney+	disney+	Twitter for iPhone
5708	1450897163790204928	"Rebelde": Netflix divulga clipe e data de lan...	papelpop	netflix	netflix	Twitter Web App
7095	1449921497720557568	@a_alinediniz @rvandromel Doces Mágnolias na N...	patricia_gomes	netflix	netflix	Twitter for iPhone
2585	1451015167685365760	eu tô tão tentado a assinar a globoplay só pra...	ItsMarchiori	globoplay	globoplay	Twitter for iPhone
4308	1452219366603362304	Para os órfãos de Rick and Morty, achei essa o...	Tiagoarruda666	netflix	netflix	Twitter for Android
947	1450658198285860864	O mc gui salvou o rico da roça kkkkkkkkk a car...	Eduardo92375039	disney+	disney+	Twitter for iPhone
3333	1451697896487849984	hbo max me salvando c ben 10 classico liga da...	afeholipe	hbo max	hbomax	Twitter for Android
1524	1449819480117465088	@calabresadani pelo amor de Deus quero morar n...	gabriela_srs	disney+	disney+	Twitter for iPhone
2759	1450992490807537664	@ccaarrvvaa Ta passando amiga, no globoplay	twitamandyy	globoplay	globoplay	Twitter for iPhone
1588	1449598443538763776	@Biaaa_comenta Pior q não, está palestrando co...	porraajess	disney+	disney+	Twitter for Android

Fonte: Elaborada pelo professor.

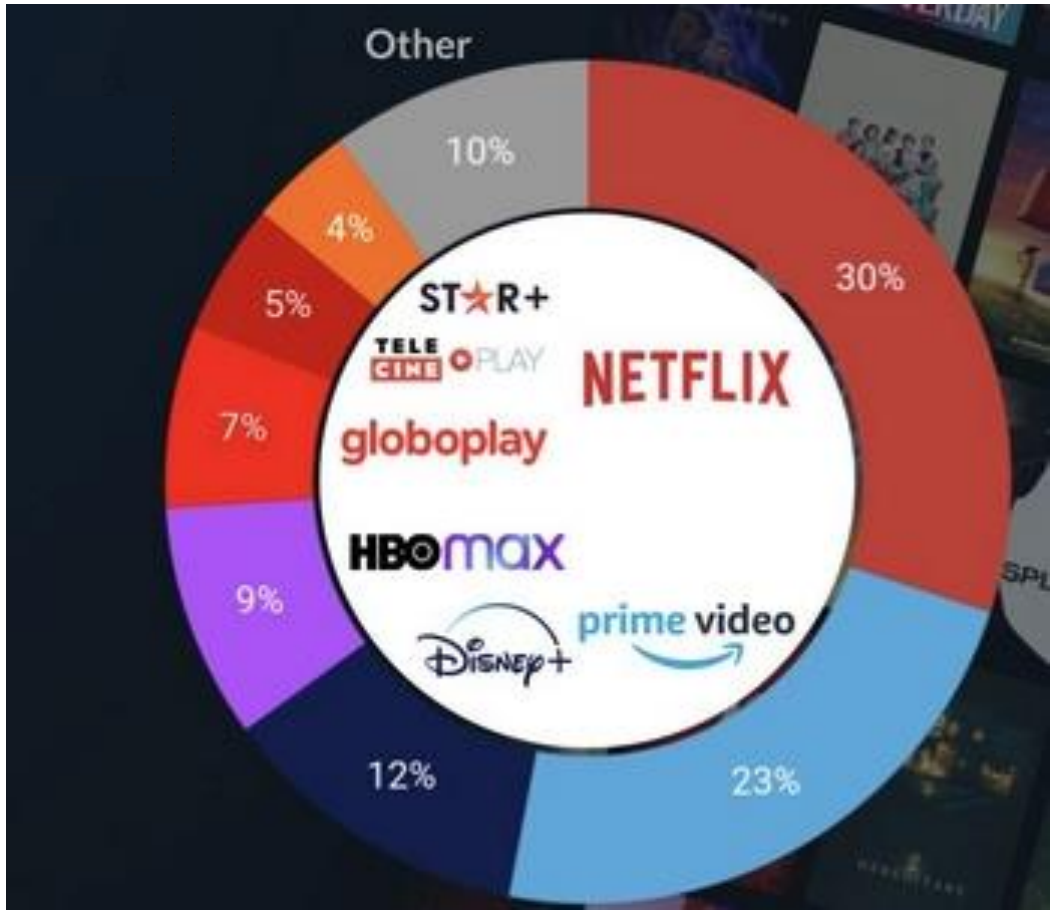
Fluxograma geral de PLN



Fonte: Elaborada pelo professor.

Caso de Uso – Guerra dos *streamings*

- Há dezenas de serviços de vídeo sob demanda disponíveis no Brasil.
- No Brasil, são cinco concorrentes principais: **Netflix**, **Prime Vídeo**, **globoplay**, **HBO MAX** (HBO GO) e **Disney+**.
- As pessoas usam o **Twitter** para falar dos serviços entre seus seguidores e emitir opiniões.



Fonte: JustWatch.com 3º trimestre 2021

IBM Cloud – Watson

- Watson é a plataforma de serviços cognitivos da IBM para negócios.
- Criado em 2010 como um sistema de computação de perguntas e respostas:
  - processamento de linguagem natural;
  - recuperação de informações;
  - representação de conhecimento;
  - raciocínio automatizado; e
  - *machine learning*.
- Disponível como serviço na IBM Cloud , inclusive em português.

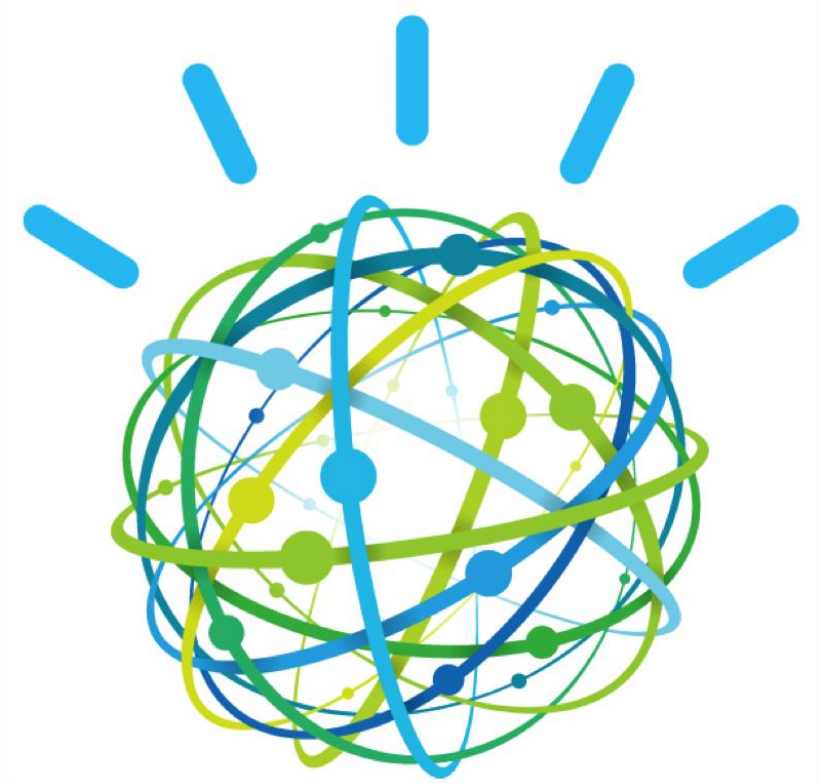
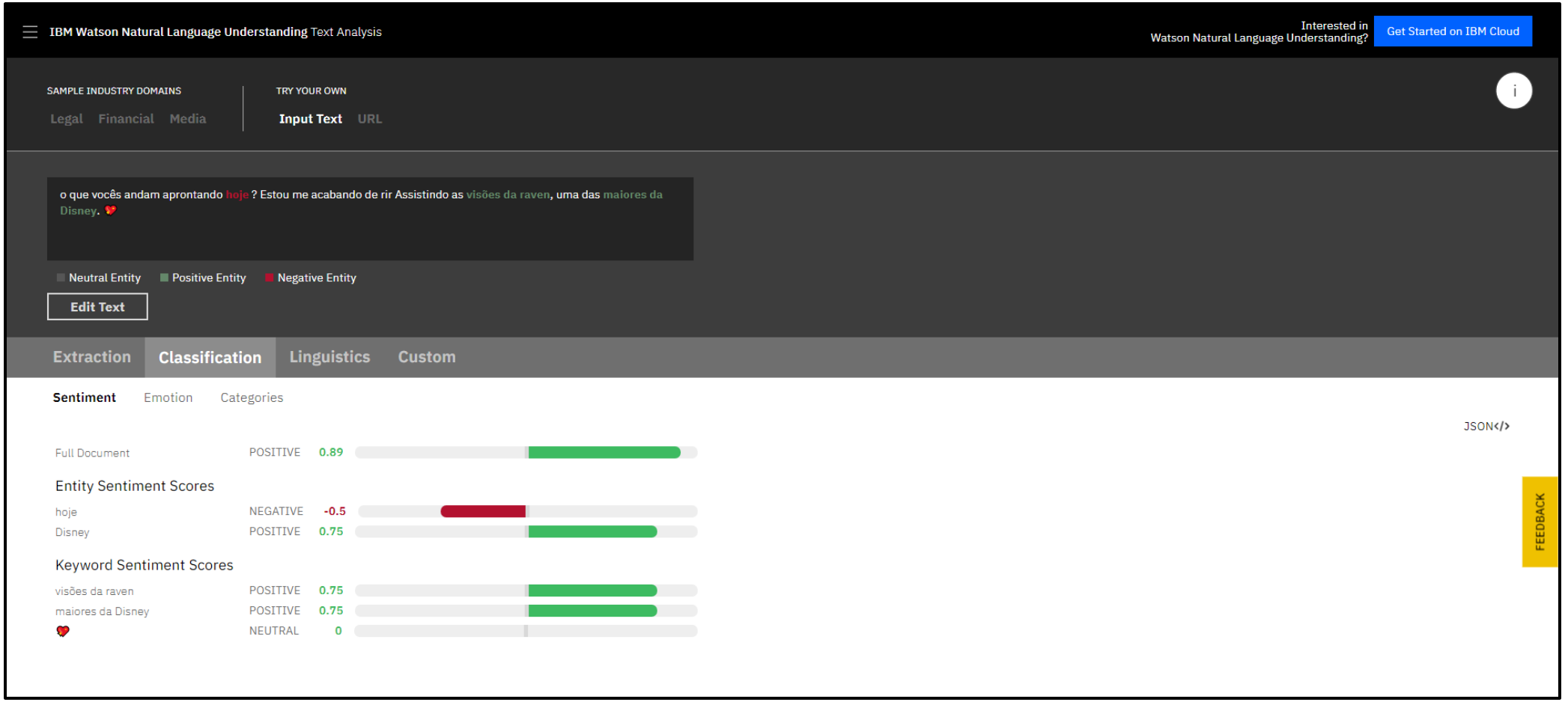


Ilustração do professor, baseada na apostila português IV da fundação CECIERJ.

IBM Cloud – Watson



Fonte: <https://www.ibm.com/demos/live/natural-language-understanding/self-service/home>.

Resultado da Análise de Sentimento com IBM Watson

Termo de busca	super negativo	negativo	levemente negativo	neutro	levemente positivo	positivo	super positivo	positivo – negativo
disney+	22%	12%	6%	24%	6%	13%	17%	-3%
globoplay	23%	14%	6%	18%	6%	12%	21%	-4%
hbomax	13%	10%	5%	32%	6%	15%	18%	10%
netflix	20%	13%	7%	18%	7%	13%	22%	3%
primevideo	9%	8%	5%	30%	6%	15%	27%	27%
Total Geral	**20%**	**12%**	**6%**	**21%**	**7%**	**13%**	**20%**	**1%**

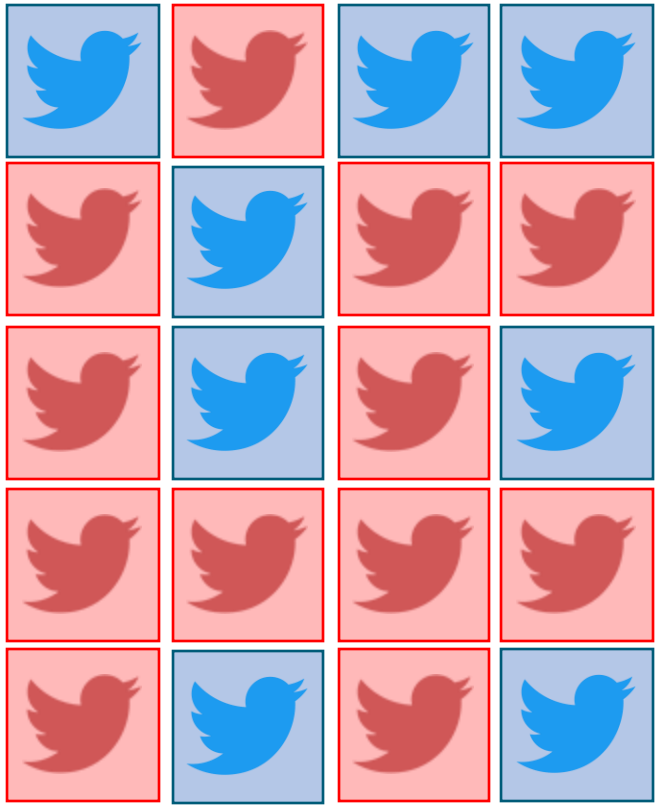
Teorema de Bayes

- Baseado nos trabalhos do pastor Thomas Bayes sobre a **teoria da probabilidade**.
- Muito utilizado em problemas de **classificação**, principalmente de textos.
- Descreve a **probabilidade de um evento** baseado em um conhecimento ***a priori***.

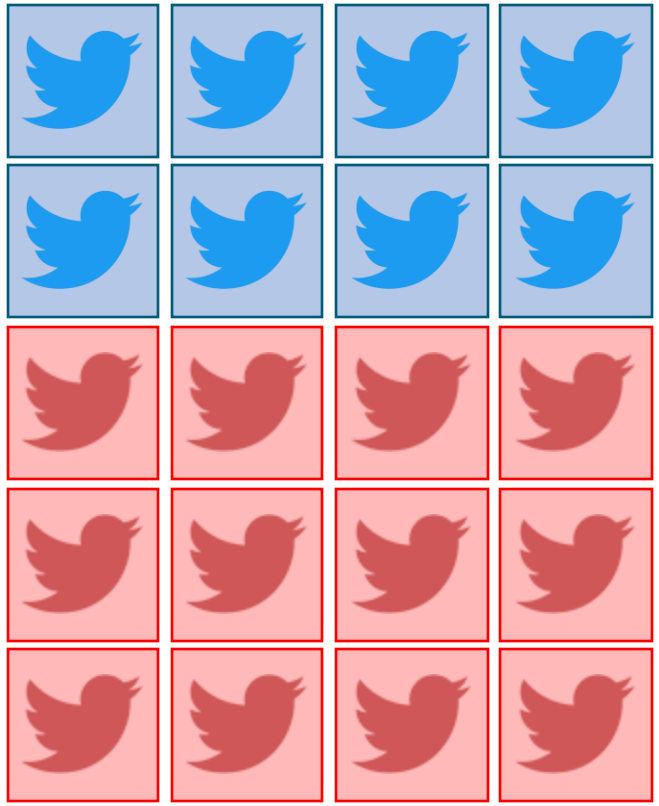
$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Thanks for Playing the Game Show Show Kevin Standlee no Flickr. – CC BY-AS 2.0

Filtro de Bayes (simplificado)



Filtro de Bayes (simplificado)



$$P(\text{"}pos\text{"}) = \frac{8}{20} = 40\%$$

$$P(\text{"}neg\text{"}) = \frac{12}{20} = 60\%$$

Filtro de Bayes (simplificado)

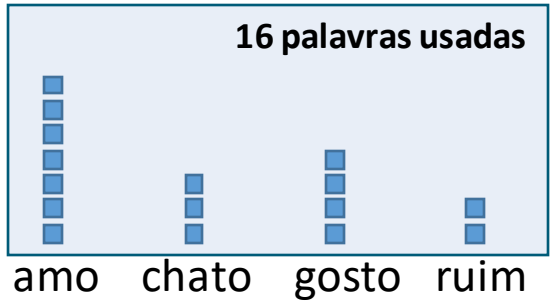
$P("pos") = 40\%$

amo chato gosto ruim

$P("neg") = 60\%$

Filtro de Bayes (simplificado)

$P("pos") = 40\%$



$$P(\text{amo}|"pos") = \frac{7}{16} = 44\%$$

$P("neg") = 60\%$

Filtro de Bayes (simplificado)

$P("pos") = 40\%$

$P(\text{amo}|"pos") = 44\%$

$$P(\text{chato}|"pos") = \frac{3}{16} = 19\%$$

$P("neg") = 60\%$

Filtro de Bayes (simplificado)

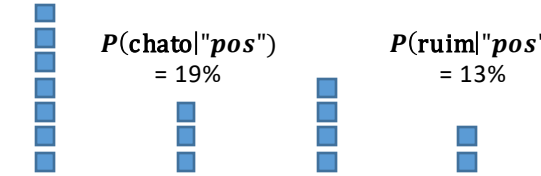
$P("pos") = 40\%$

$P(\text{amo}|"pos") = 44\%$

$P(\text{gosto}|"pos") = 25\%$

$P(\text{chato}|"pos") = 19\%$

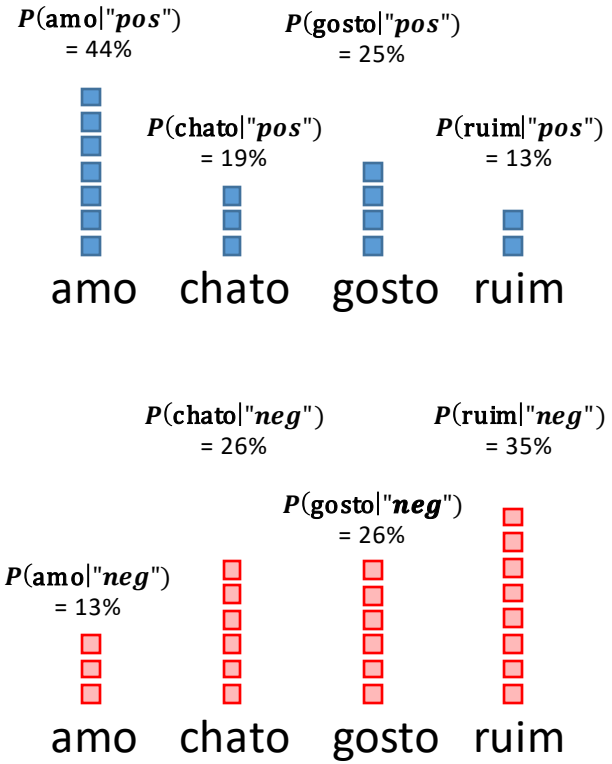
$P(\text{ruim}|"pos") = 13\%$



$P("neg") = 60\%$

Filtro de Bayes (simplificado)

$P(\text{"}pos\text{"}) = 40\%$



$P(\text{"}neg\text{"}) = 60\%$

Filtro de Bayes (simplificado)

$P("pos") = 40\%$

$P(amo|"pos") = 44\%$

$P(chato|"pos") = 19\%$

$P(gosto|"pos") = 25\%$

$P(ruim|"pos") = 13\%$

$P(amo|"neg") = 13\%$

$P(chato|"neg") = 26\%$

$P(gosto|"neg") = 26\%$

$P(ruim|"neg") = 35\%$

$P("neg") = 60\%$

"Eu **amo** a HBO MAX"

$$P("pos" \mid amo) = \frac{P(amo \mid "pos") \cdot P("pos")}{P(amo)}$$

$$P("neg" \mid amo) = \frac{P(amo \mid "neg") \cdot P("neg")}{P(amo)}$$

Filtro de Bayes (simplificado)

$P("pos") = 40\%$

$P(\text{amo}|"pos") = 44\%$

$P(\text{chato}|"pos") = 19\%$

$P(\text{gosto}|"pos") = 25\%$

$P(\text{ruim}|"pos") = 13\%$

$P(\text{amo}|"neg") = 13\%$

$P(\text{chato}|"neg") = 26\%$

$P(\text{gosto}|"neg") = 26\%$

$P(\text{ruim}|"neg") = 35\%$

$P("neg") = 60\%$

"Eu **amo** a HBO MAX"

$$P("pos" \mid amo) \propto P(amo \mid "pos") \cdot P("pos")$$

$$P("neg" \mid amo) \propto P(amo \mid "neg") \cdot P("neg")$$

Filtro de Bayes (simplificado)

$P("pos") = 40\%$

$P(\text{amo}|"pos") = 44\%$

$P(\text{chato}|"pos") = 19\%$

$P(\text{gosto}|"pos") = 25\%$

$P(\text{ruim}|"pos") = 13\%$

$P(\text{amo}|"neg") = 13\%$

$P(\text{chato}|"neg") = 26\%$

$P(\text{gosto}|"neg") = 26\%$

$P(\text{ruim}|"neg") = 35\%$

$P("neg") = 60\%$

"Eu **amo** a HBO MAX"

$$P("pos" \mid amo) \propto P(amo \mid "pos") \cdot 40\%$$

$$P("neg" \mid amo) \propto P(amo \mid "neg") \cdot P("neg")$$

Filtro de Bayes (simplificado)

$P("pos") = 40\%$

$P(\text{amo}|"pos") = 44\%$

$P(\text{chato}|"pos") = 19\%$

$P(\text{gosto}|"pos") = 25\%$

$P(\text{ruim}|"pos") = 13\%$

$P(\text{amo}|"neg") = 13\%$

$P(\text{chato}|"neg") = 26\%$

$P(\text{gosto}|"neg") = 26\%$

$P(\text{ruim}|"neg") = 35\%$

$P("neg") = 60\%$

"Eu **amo** a HBO MAX"

$$P("pos" \mid amo) \propto 44\% \cdot 40\%$$

$$P("neg" \mid amo) \propto P(amo \mid "neg") \cdot P("neg")$$

Filtro de Bayes (simplificado)

$P("pos") = 40\%$

$P(\text{amo}|"pos") = 44\%$

$P(\text{chato}|"pos") = 19\%$

$P(\text{gosto}|"pos") = 25\%$

$P(\text{ruim}|"pos") = 13\%$

$P(\text{amo}|"neg") = 13\%$

$P(\text{chato}|"neg") = 26\%$

$P(\text{gosto}|"neg") = 26\%$

$P(\text{ruim}|"neg") = 35\%$

$P("neg") = 60\%$

"Eu **amo** a HBO MAX"

$$P("pos" \mid amo) \propto 44\% \cdot 40\% = 17,6\%$$

$$P("neg" \mid amo) \propto P(amo \mid "neg") \cdot P("neg")$$

Filtro de Bayes (simplificado)

$P(\text{"}pos\text{"}) = 40\%$

$P(\text{amo}|\text{"}pos\text{"})$ = 44%

$P(\text{chato}|\text{"}pos\text{"})$ = 19%

$P(\text{gosto}|\text{"}pos\text{"})$ = 25%

$P(\text{ruim}|\text{"}pos\text{"})$ = 13%

$P(\text{amo}|\text{"}neg\text{"})$ = 13%

$P(\text{chato}|\text{"}neg\text{"})$ = 26%

$P(\text{gosto}|\text{"}neg\text{"})$ = 26%

$P(\text{ruim}|\text{"}neg\text{"})$ = 35%

$P(\text{"}neg\text{"}) = 60\%$

"Eu **amo** a HBO MAX"

$$P(\text{"}pos\text{"} \mid amo) \propto 44\% \cdot 40\% = 17,6\%$$

$$P(\text{"}neg\text{"} \mid amo) \propto 13\% \cdot 60\%$$

Filtro de Bayes (simplificado)

$P(\text{"}pos\text{"}) = 40\%$

$P(\text{amo}|\text{"}pos\text{"})$ = 44%

$P(\text{chato}|\text{"}pos\text{"})$ = 19%

$P(\text{gosto}|\text{"}pos\text{"})$ = 25%

$P(\text{ruim}|\text{"}pos\text{"})$ = 13%

$P(\text{amo}|\text{"}neg\text{"})$ = 13%

$P(\text{chato}|\text{"}neg\text{"})$ = 26%

$P(\text{gosto}|\text{"}neg\text{"})$ = 26%

$P(\text{ruim}|\text{"}neg\text{"})$ = 35%

$P(\text{"}neg\text{"}) = 60\%$

"Eu **amo** a HBO MAX"

$$P(\text{"}pos\text{"} \mid amo) \propto 44\% \cdot 40\% = 17,6\%$$

$$P(\text{"}neg\text{"} \mid amo) \propto 13\% \cdot 60\% = 7,8\%$$

Filtro de Bayes (simplificado)

$P("pos") = 40\%$

$P(\text{amo}|"pos") = 44\%$

$P(\text{chato}|"pos") = 19\%$

$P(\text{gosto}|"pos") = 25\%$

$P(\text{ruim}|"pos") = 13\%$

$P(\text{amo}|"neg") = 13\%$

$P(\text{chato}|"neg") = 26\%$

$P(\text{gosto}|"neg") = 26\%$

$P(\text{ruim}|"neg") = 35\%$

$P("neg") = 60\%$

"Eu **amo** a HBO MAX"

Positiva

Código

```
1.  # Importa as bibliotecas
2.  ...
3.
4.  # carrega as amostras de Tweet para treinamento e remove as neutras
5.  df_nbr = pd.read_csv('amostras-revisado.csv', delimiter=';')
6.  df_nbr = df_nbr.loc[df_nbr['sentiment_revised'].ne(0)]
7.
8.  # faz o pre-tratamento
9.  ...
10.
11. # Separa as palavras dos textos
12. df_nbr['words'] = df_nbr['text_treated'].apply(lambda x: word_tokenize(x))
13.
14. # Faz a contagem de palavras
15. all_words = nltk.FreqDist([word.lower() for words in df_nbr['words'].to_list() for word in words])
16.
17. # Transforma em lista, para usar a biblioteca NTLK
18. words = df_nbr['words'].to_list()
19. sentiments = df_nbr['sentiment_label_revised'].to_list()
20.
21. # Prepara os documentos, no formato do NLTK (tuplas)
22. documents = [(words[i], sentiments[i]) for i, sentiment in enumerate(sentiments)]
23.
24. # Extrai as características do texto para treinamento do modelo
25. def find_features(document):
26.     words = set(document)
27.     features = {}
28.     for w in word_features:
29.         features[w] = (w in words)
30.     return features
31. featuresets = [(find_features(rev), category) for (rev, category) in documents]
32.
33. # Separa 150 registros para treinamento
34. training_set = featuresets[:150]
35. classifier = nltk.NaiveBayesClassifier.train(training_set)
36.
37. # Testa o modelo nos registros que havíamos separados para testes
38. testing_set = featuresets[150:]
39. print("Acurácia:",(nltk.classify.accuracy(classifier, testing_set))*100)
```

Resultados

```
[21]  1  classifier.show_most_informative_features(20)
```

```
Most Informative Features
                   serie = True              pos : neg    =      6.6 : 1.0
                   todos = True              pos : neg    =      4.6 : 1.0
                  filmes = True              pos : neg    =      4.6 : 1.0
                     ter = True              neg : pos    =      3.7 : 1.0
                       & = True              pos : neg    =      3.6 : 1.0
                     boa = True              pos : neg    =      3.6 : 1.0
                  melhor = True              pos : neg    =      3.6 : 1.0
                      ta = True              pos : neg    =      3.4 : 1.0
                     vai = True              neg : pos    =      3.0 : 1.0
                       ? = True              neg : pos    =      2.7 : 1.0
                  disney = True              neg : pos    =      2.7 : 1.0
                       ! = True              pos : neg    =      2.6 : 1.0
                       n = True              pos : neg    =      2.6 : 1.0
                    casa = True              pos : neg    =      2.6 : 1.0
                     obg = True              pos : neg    =      2.6 : 1.0
                   tempo = True              pos : neg    =      2.6 : 1.0
                       ' = True              pos : neg    =      2.6 : 1.0
                     you = True              pos : neg    =      2.6 : 1.0
                       a = True              pos : neg    =      2.6 : 1.0
                  disney+ = True             pos : neg    =      2.6 : 1.0
```

Universidade Presbiteriana
Mackenzie