

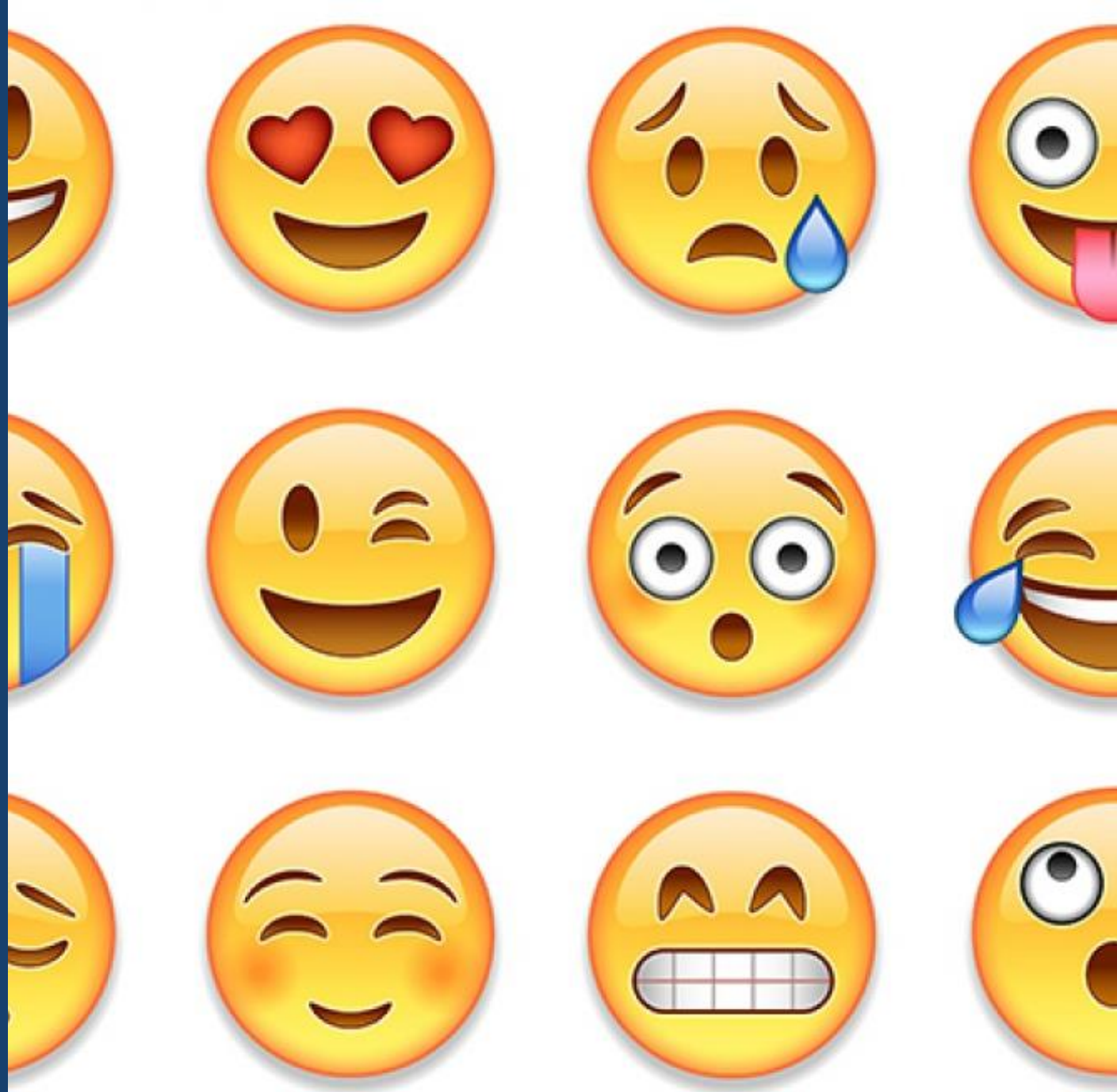
# CORONAVIRUS – ANÁLISE DE SENTIMENTOS EM REDES SOCIAIS

Profª Manoela Kohler

# Coronavirus – Análise de sentimentos em redes sociais

- Análise de sentimentos: entendendo o problema
- Biblioteca AFFIN para análise de sentimentos
- Biblioteca NLTK para análise de sentimentos
- Abordagem de Inteligência Artificial para análise de sentimentos
- Cases
- Hands on!!!

# Análise de Sentimentos: entendendo o problema



# Análise de Sentimentos: entendendo o problema

- A análise de sentimentos é a interpretação e classificação de emoções (positivas, negativas e neutras) nos dados de texto usando técnicas de análise de texto.
- A análise de sentimentos permite que as empresas identifiquem os sentimentos dos clientes em relação a produtos, marcas ou serviços em conversas e comentários on-line.
- Ao analisar automaticamente o feedback do cliente, desde as respostas da pesquisa às conversas nas mídias sociais, as marcas podem ouvir atentamente seus clientes e personalizar produtos e serviços para atender às suas necessidades.

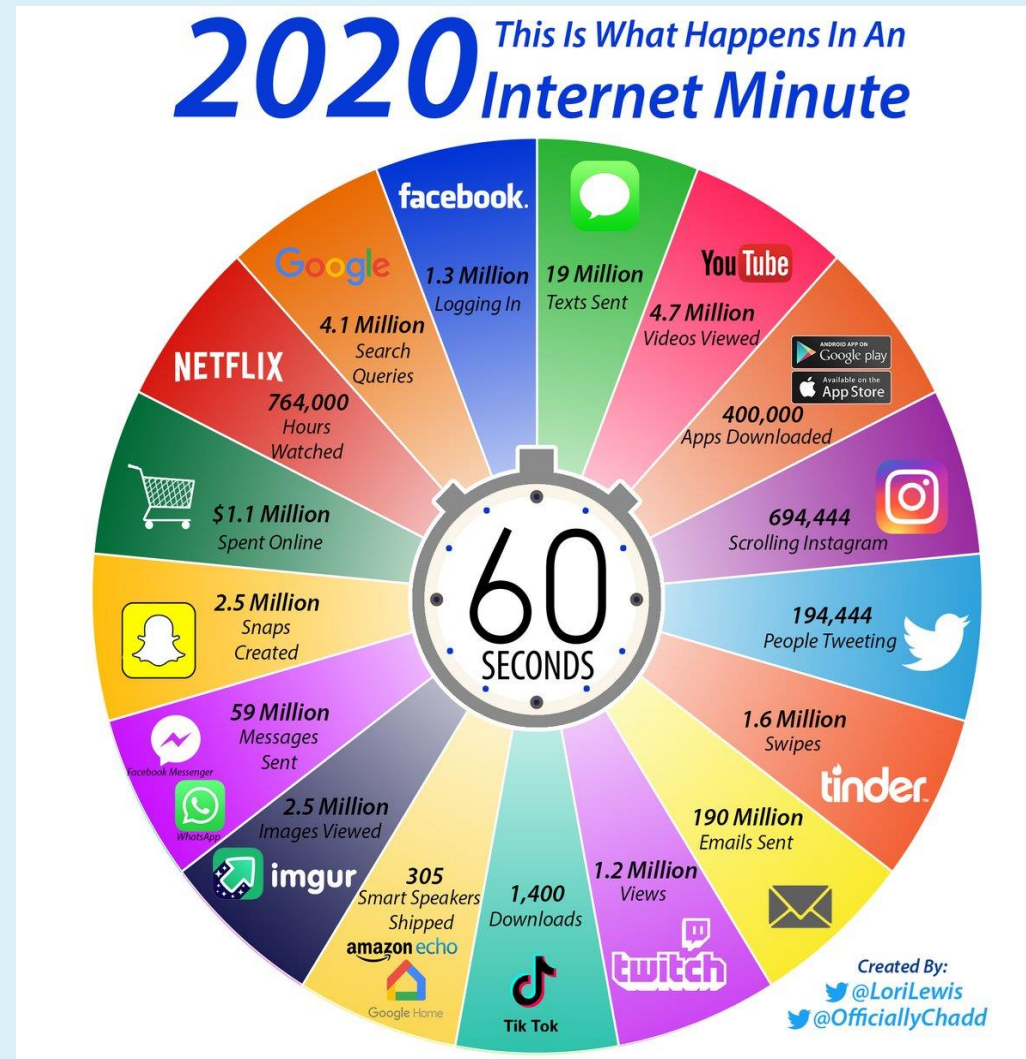
# Tipos de análise de sentimentos

- Os modelos de análise de sentimentos se concentram na polaridade (positiva, negativa, neutra), mas também nos sentimentos e emoções (raiva, alegria, tristeza, etc.) e até mesmo nas intenções (por exemplo, interessado x não interessado).
- Tipos:
  - *Análise de sentimentos refinado (refinar além de positivo, negativo e neutro);*
  - *Detecção de emoção (raiva, tristeza, alegria, etc);*
  - *Análise de sentimentos baseada em aspectos (de um produto, por exemplo);*
- Modelo:
  - *Léxico (repositório de palavras e suas emoções);*
  - *Modelos complexos de machine learning.*

# Por que fazer análise de sentimentos

- Estima-se que 80% dos dados do mundo não sejam estruturados:
  - *Um grande volume de dados de texto (e-mails, tíquetes de suporte, bate-papos, conversas em mídias sociais, pesquisas, artigos, documentos, etc.) é criado todos os dias, mas é difícil de analisar, entender e classificar.*
- A análise de sentimentos, no entanto, ajuda as empresas a entender todo esse texto não estruturado, marcando-o automaticamente.

# Por que fazer análise de sentimentos



# Benefícios

- Classificação de grandes conjuntos de dados de maneira eficiente e econômica.
- Identificação de problemas críticos em tempo real para tomada de decisão imediata.
- Critérios consistentes para classificação dos sentimentos (já que classificação feita por pessoas diferentes pode ser altamente subjetiva), ajudando a melhorar a precisão e obter melhores insights.



# Préprocessamento de texto

- Tokenização
- Remoção de *urls*
- *Remoção de caracteres especiais e acentos*
- Remoção de *stopwords*
  - *Remoção de palavras com baixo valor de discriminação para o processo de recuperação da informação.*
- *Stemming e Lematização*
  - *O objetivo da lematização e do stemming é reduzir (deflexionar) uma palavra a sua base. A diferença entre os dois é que o stemming corta a palavra tentando acertar a sua base na maioria da vezes, enquanto a lematização reduz a base utilizando um vocabulário e a análise morfológica das palavras.*

# Préprocessamento de texto

## ■ Stemming e Lematização

*Exemplos:*

- 1. A palavra "walk" é a base para "walking", e é corretamente reduzida pelo stemming e pela lematização.*
- 2. A palavra "better" tem "good" como base (ou lema). Esse link é perdido pelo stemming.*
- 3. A palavra "meeting" pode ser reduzida para um substantivo ou um verbo dependendo do contexto.*

*E.g., "in our last meeting" ou "We are meeting again tomorrow".*

*A lematização consegue fazer a redução de forma correta.*

# Modelos para análise de sentimentos

- Sistemas baseados em regras
- *Machine Learning*
- Sistemas híbridos

# Modelos para análise de sentimentos

- Sistemas baseados em regras
- *Machine Learning*
- Sistemas híbridos

# Sistemas baseados em regras

## ■ Exemplo:

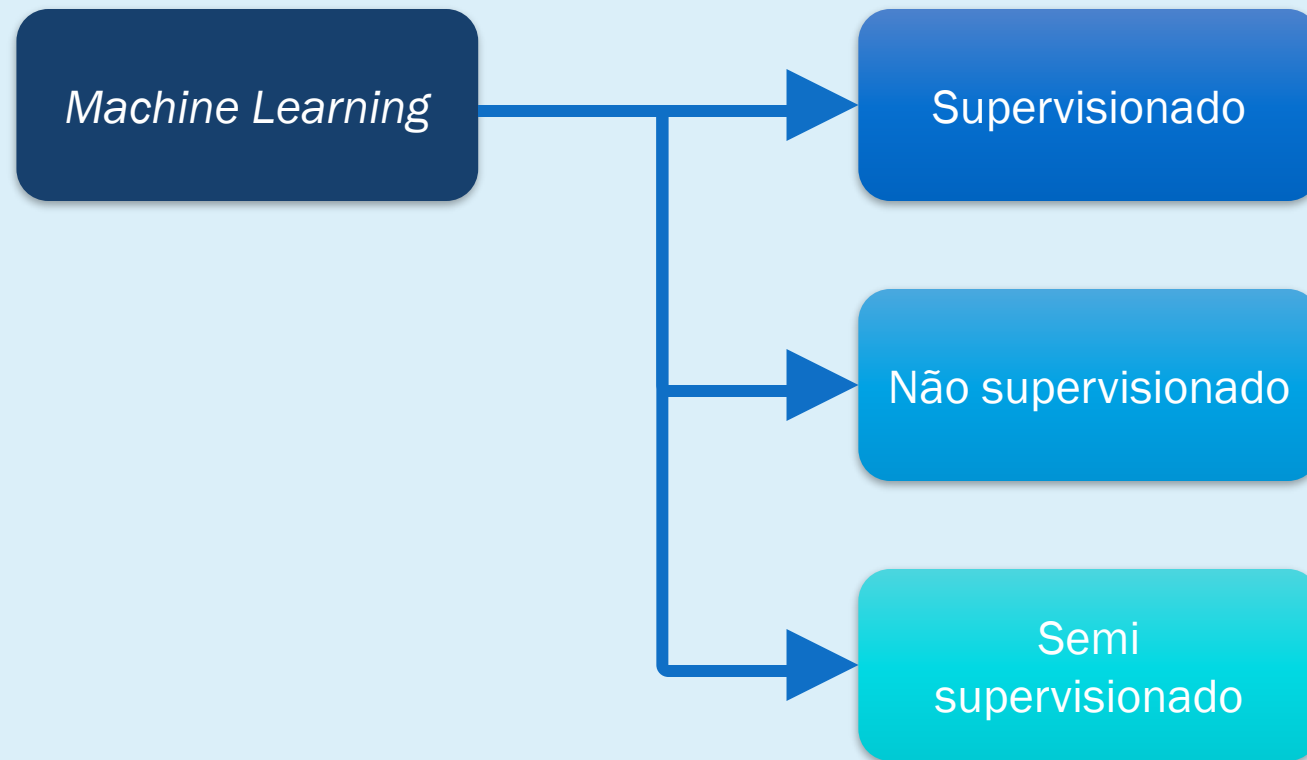
1. Define duas listas de palavras polarizadas (palavras positivas e palavras negativas)
2. Conta-se o número de palavras positivas e negativas que aparecem em um determinado texto
3. Se o número de aparições de palavras positivas for maior que o número de aparências de palavras negativas, o sistema retornará um sentimento positivo e vice-versa. Se for igual, o sistema retornará um sentimento neutro.

Modelo ingênuo;  
Pode se tornar muito  
complexo e de difícil  
manutenção.

# Modelos para análise de sentimentos

- Sistemas baseados em regras
- *Machine Learning*
- Sistemas híbridos

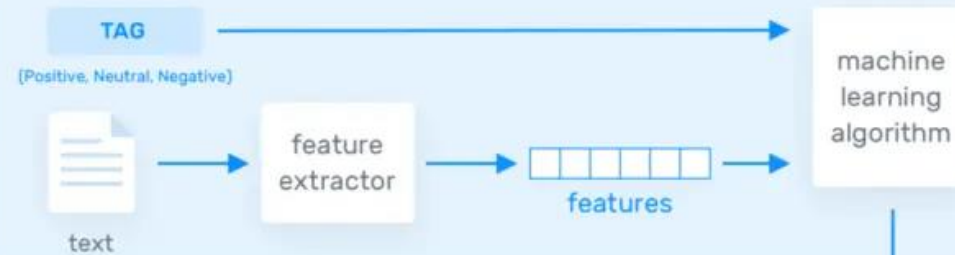
# Machine Learning



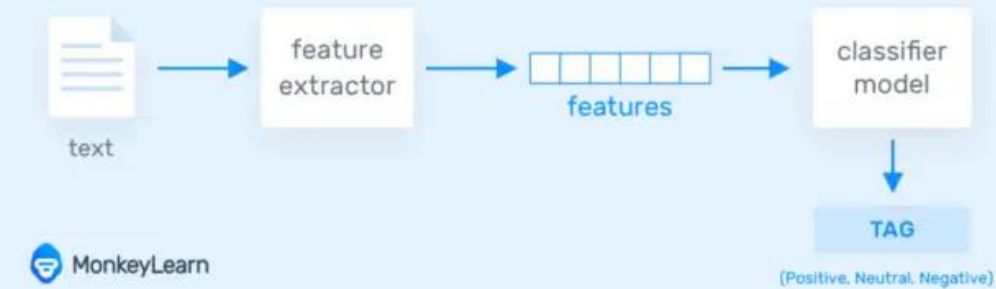
# Machine Learning

## How Does Sentiment Analysis Work?

### (a) Training

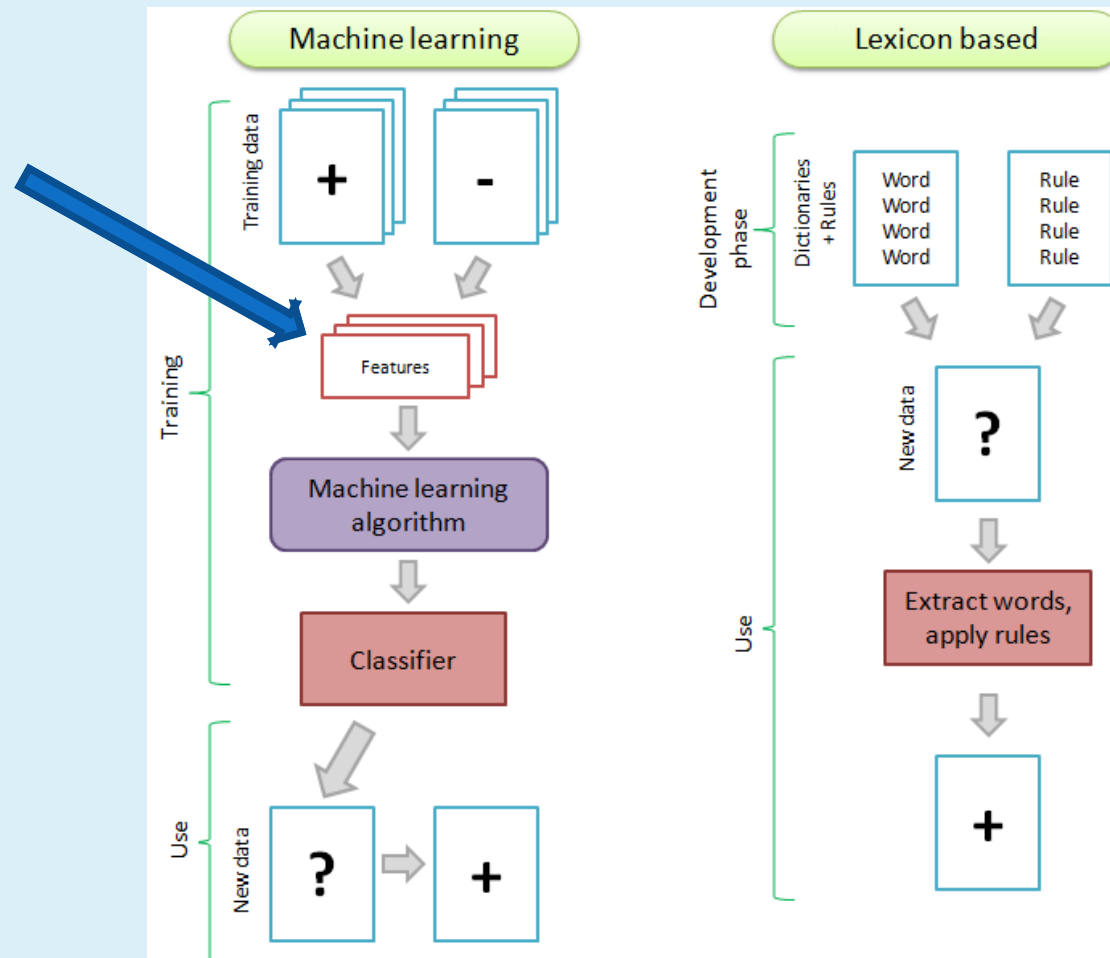


### (b) Prediction





# Machine Learning



# Extração de características

Transformação de palavras, sentenças ou documentos em vetores numéricos:

		Dimensions					
Word vectors	dog	-0.4	0.37	0.02	-0.34	animal	
	cat	-0.15	-0.02	-0.23	-0.23	domesticated	
	lion	0.19	-0.4	0.35	-0.48	pet	
	tiger	-0.08	0.31	0.56	0.07	fluffy	
	elephant	-0.04	-0.09	0.11	-0.06		
	cheetah	0.27	-0.28	-0.2	-0.43		
	monkey	-0.02	-0.67	-0.21	-0.48		
	rabbit	-0.04	-0.3	-0.18	-0.47		
	mouse	0.09	-0.46	-0.35	-0.24		
	rat	0.21	-0.48	-0.56	-0.37		

# Machine Learning – Processamento de Linguagem Natural

## ■ Extração de características ou vetorização de palavras:

- *Bag of Words (BoW)*

Documento	Eu	amo	cachorro	gato
Eu amo cachorro	1	1	1	0
Eu amo gato	1	1	0	1

- *Perde-se ordenação das palavras*
- *Vetor pode se tornar muito grande*
- *Representação esparsa*
- *Na inferência, palavras que não estavam no corpus são ignoradas*
- *Simples e independente da língua*
- *N-gram*

# Machine Learning – Processamento de Linguagem Natural

- Extração de características ou vetorização de palavras:

- *Contagem de palavras e Frequência de palavras*

Documento	Eu	amo	cachorro	e	gato
Eu amo cachorro e amo gato	1	2	1	1	1
Eu amo gato	1	1	0	0	1

- *TFIDF (Term Frequency – Inverse Document Frequency)*

- Resolve problemas dos métodos de contagem e frequência, onde palavras que aparecem muitas vezes, acabam dominando os documentos, mas possivelmente contém informação que não é tão importante;
    - Re-escala a frequência de forma a penalizar palavras que são muito frequentes em todos os documentos, como artigos, por exemplo.

# Machine Learning – Processamento de Linguagem Natural

## ■ Word2Vec

- Mikolov, 2013 - Google (<https://code.google.com/archive/p/word2vec/>)
- Modelo pré-treinado em um dataset de do Google News (100 bilhões de palavras)
- Modelo aprende a representar palavras através de vetores, capturando regularidades linguísticas:

### France: Word Cosine distance

spain	0.678515
belgium	0.665923
netherlands	0.652428
italy	0.633130
switzerland	0.622323
luxembourg	0.610033
portugal	0.577154
russia	0.571507
germany	0.563291
catalonia	0.534176

## word2Vec

King – Man + Woman  $\approx$  Queen

Paris – France + Italy  $\approx$  Rome

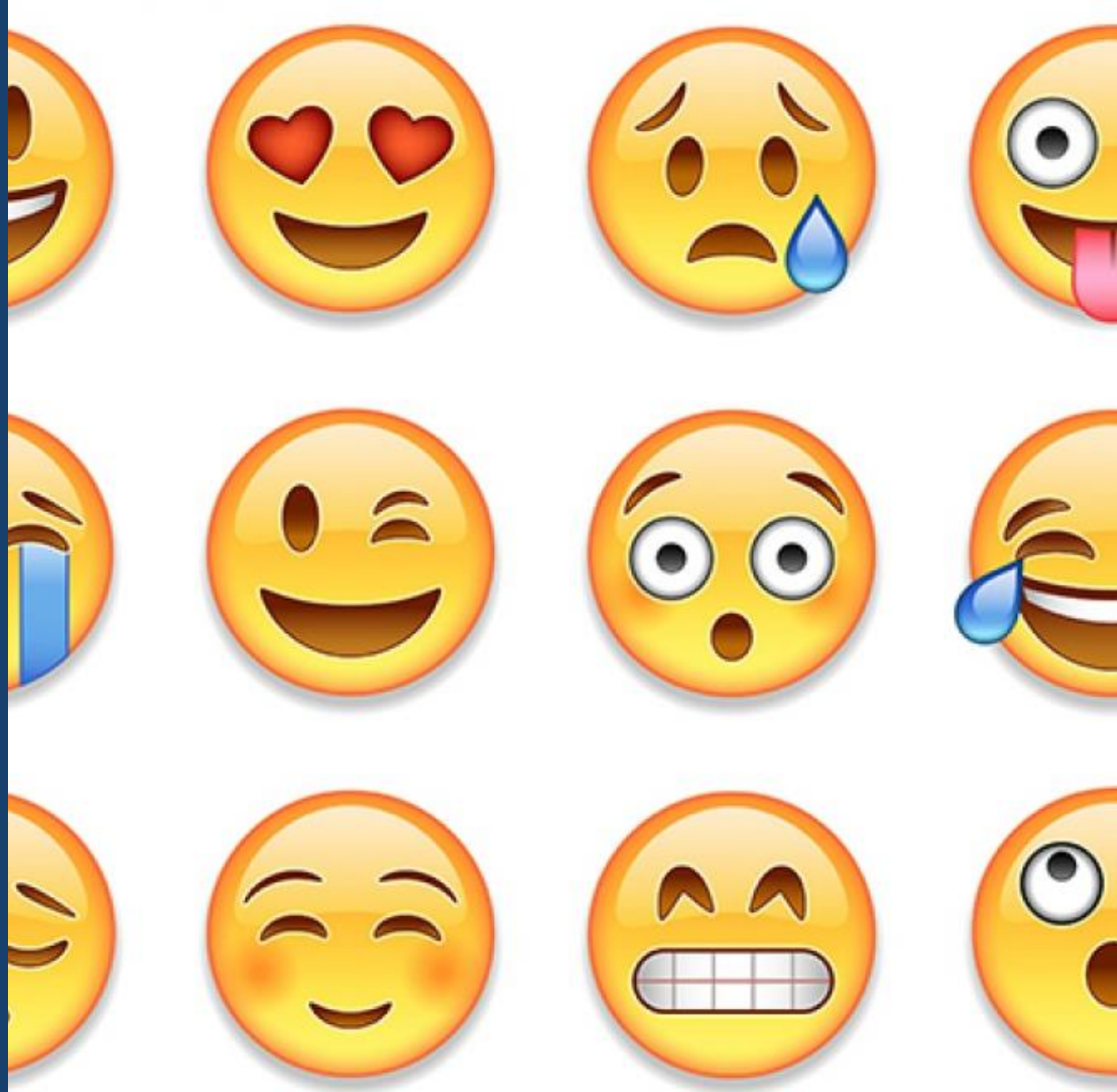
# Machine Learning – Processamento de Linguagem Natural

- Word2vec: polissemia    **‘GO’ pode ser um verbo, mas também um jogo de tabuleiro!**



- Sentence to Vector:
  - BERT (Google, 2018; <https://arxiv.org/abs/1810.04805>)
  - RoBERTa (Facebook, 2019; <https://arxiv.org/abs/1907.11692>)

Biblioteca  
AFFIN para  
análise de  
sentimentos



■ AFFIN (Finn Arup Nielsen, 2014)

- *Ferramenta de análise de sentimentos baseada em léxico e regras.*
- *Especificamente construída para trabalhar com microblogs.*

<i>poor</i>	<i>-1</i>
<i>bad</i>	<i>-2</i>
<i>terrible</i>	<i>-5</i>
<i>good</i>	<i>+1</i>
<i>great</i>	<i>+4</i>

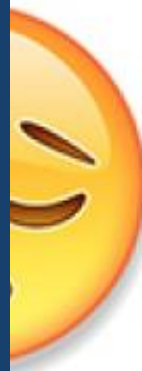
$$\text{Score\_doc} = \frac{1}{n} \sum_{p=1}^n \text{score}_p$$

Onde  $n$  é o número de palavras do documento.

<http://www2.imm.dtu.dk/pubdb/edoc/imm6006.pdf>

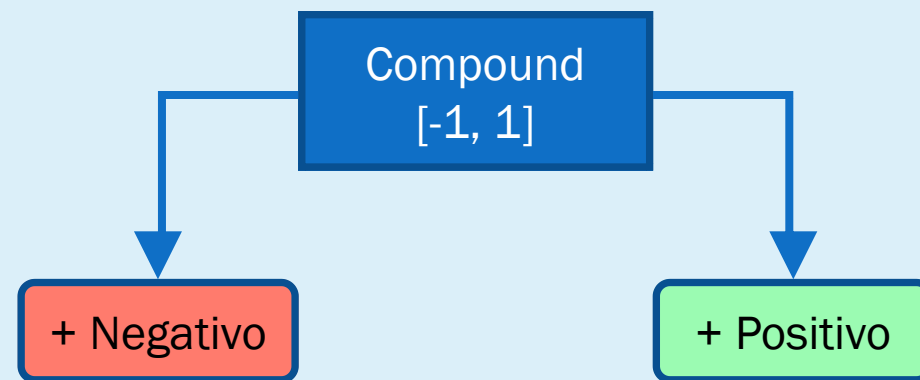


# Biblioteca NLTK para análise de sentimentos



# Biblioteca NLTK para análise de sentimentos

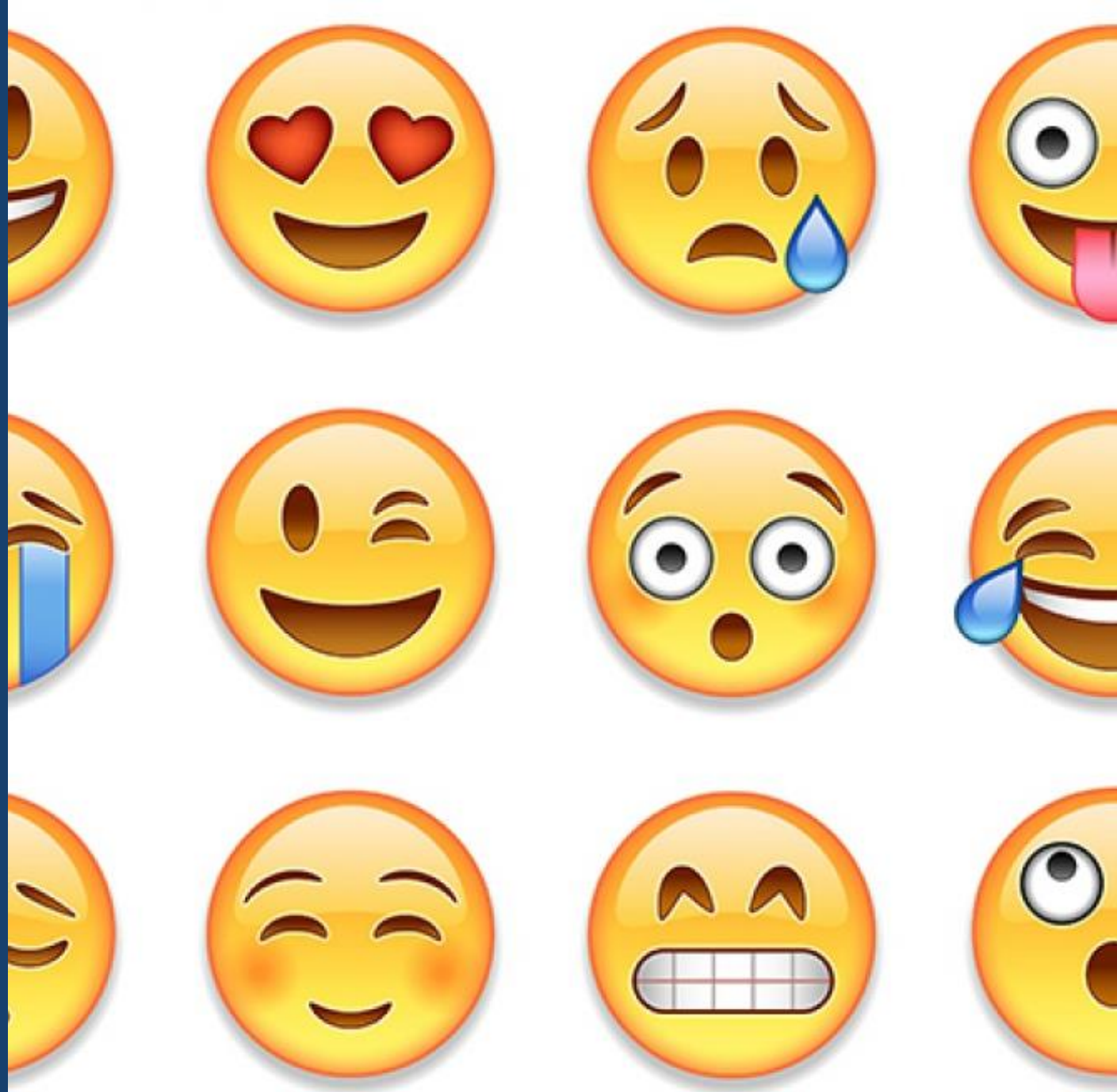
- Vader (Valence Aware Dictionary and sEntiment Reasoner, 2014)
  - *Ferramenta de análise de sentimentos, recentemente incorporado a biblioteca NLTK, baseada em léxico e regras.*
  - *Especificamente construída para trabalhar com textos de redes sociais.*



\*\* Normalizado entre -1 e 1

<https://github.com/cjhutto/vaderSentiment>

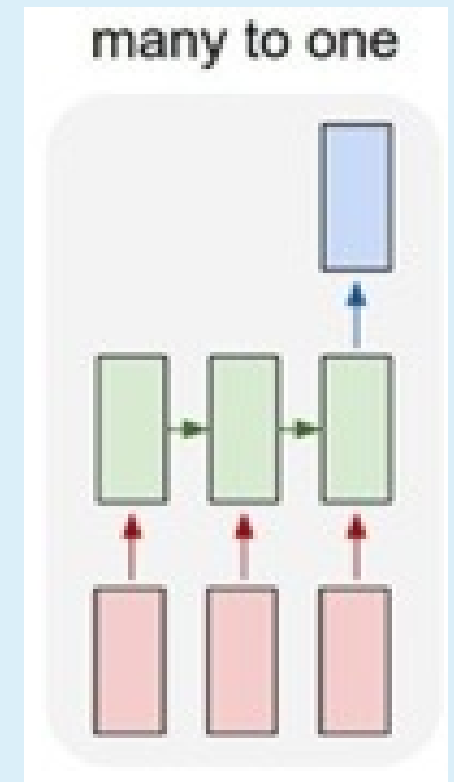
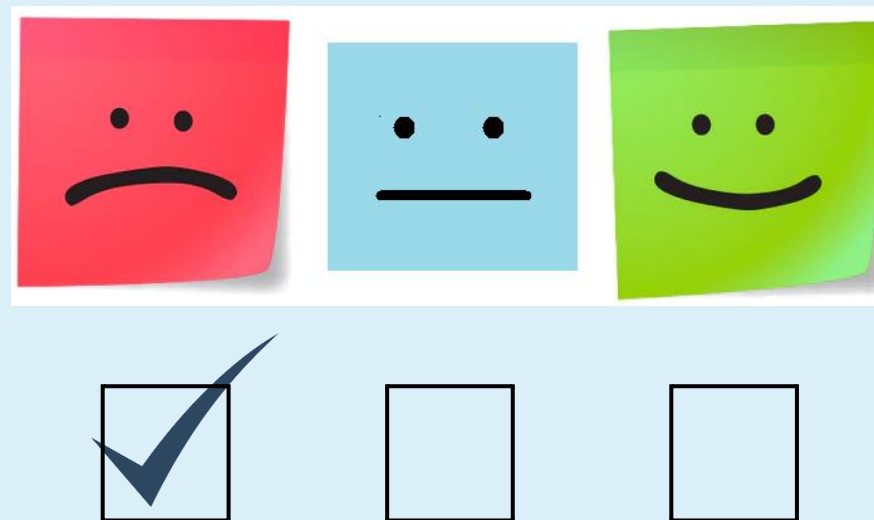
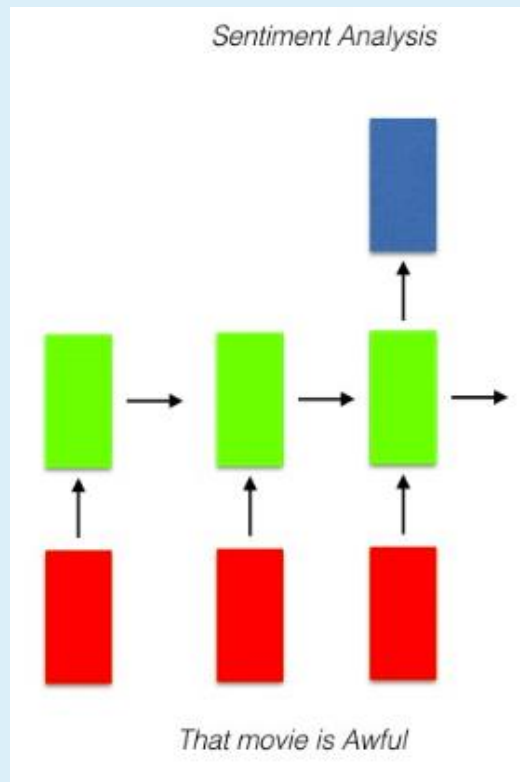
# Abordagem de Inteligência Artificial para análise de sentimentos



# Redes Recorrentes: arquitetura Many to One

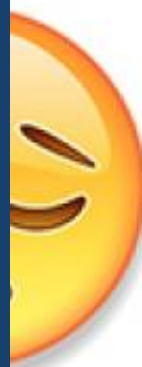
**Entrada:** Sequência de palavras

**Saída:** Classificar se o sentimento é positivo ou negativo ou neutro





# Cases



# Cases

## ■ Eleições presidenciais dos EUA, 2016 (<https://aisel.aisnet.org/pacis2017/48>)

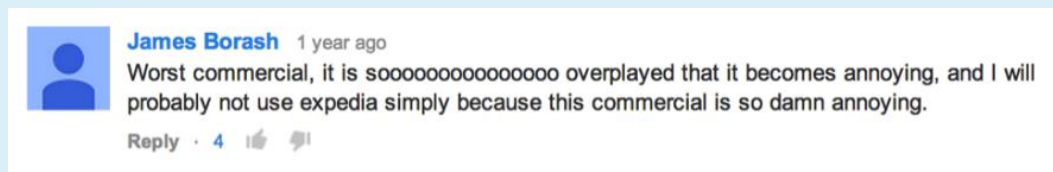
- *Volume de tweets sobre Trump foram muito maiores que sobre Clinton;*
- *A propagação do negativismo foi muito mais forte em direção a Clinton (causando sentimentos de desconfiança), apesar de ambos terem muito mais tweets negativo que positivos;*
- *Trump soube usar o Twitter para alcançar seu público alvo longe da mídia tradicional.*



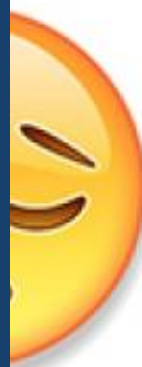
# Cases

## ■ Expedia Canada (2015)

- *Após o lançamento de um novo comercial, a empresa percebeu que a grande maioria dos tweets sobre o comercial eram negativos;*
- *A Expedia pode atuar a tempo, mudando a música do comercial e fazendo uma brincadeira em relação ao som do violino que causou tanta irritação às pessoas, chamando mais atenção (positiva) nas redes sociais.*



Hands on!!!





# Hands on!!!

- Nossa prática será dividida em duas partes.
  - *Parte 1: Baixar do Twitter um conjunto de tweets sobre o tema de interesse: **coronavirus**.*
  - *Parte 2: Fazer a análise de sentimentos nos tweets baixados e fazer algumas análises gráficas, como wordcloud, boxplots, histogramas, etc.*

*\*Os dois scripts foram criados de forma a poderem ser facilmente reutilizados para outras análises no Twitter.*