



# PROJETO - TEXT MINING

---

Grupo 7



# ÍNDICE

Revisão de literatura

Descrição dos Dados e Métrica de Avaliação

Baseline de Ferramentas Existentes

Aplicação de Léxico de Sentimentos

Modelo Aprendizagem Automática

Modelo Transformadores

Modelos Generativos

Conclusão e Discussão

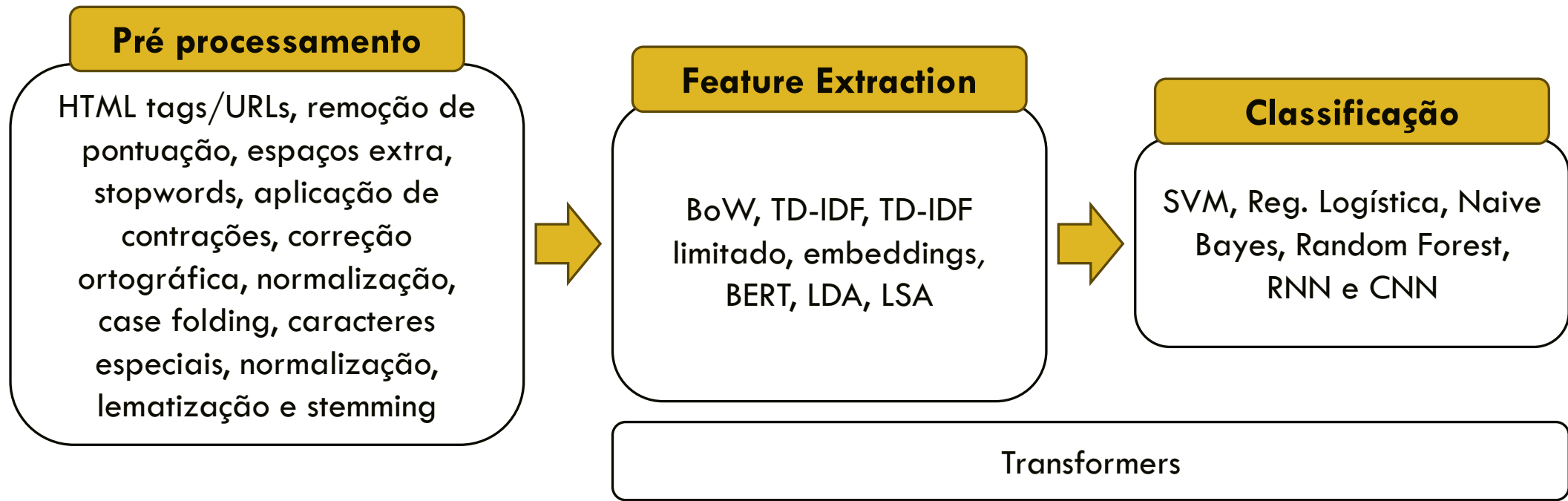
Propostas para Trabalho Futuro

Anexos



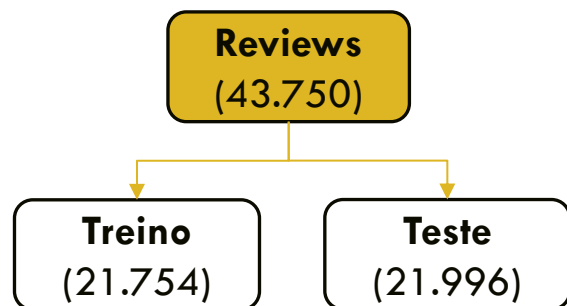
# REVISÃO DE LITERATURA

- Durante a revisão de literatura, observou-se que a combinação de diferentes técnicas de pré-processamento, extração de features e modelos de classificação é determinante para o desempenho na análise de sentimento do IMDb



- ML: 86%-91% e Transformers 94%-98%**

## DESCRIÇÃO DOS DADOS E MÉTRICA DE AVALIAÇÃO

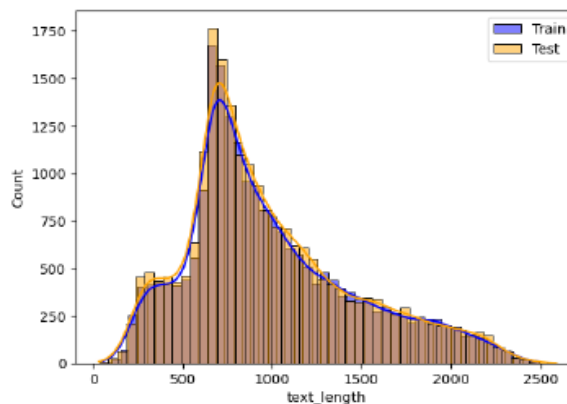


- 1% de diferença (ou 242 reviews), assumimos balanceado
- As categorias negativas e positivas também estavam balanceadas em ambos os conjuntos

## Análise de bigramas (pos e neg)

Positivo			Negativo		
	bigram	count		bigram	count
0	ve seen	710	0	ve seen	944
1	good movie	336	1	waste time	687
2	don know	320	2	special effects	536
3	special effects	313	3	don know	463
4	new york	299	4	low budget	457
5	great movie	297	5	movie just	437
6	10 10	293	6	looks like	435
7	years ago	285	7	bad movie	385
8	saw movie	267	8	worst movie	349
9	real life	264	9	year old	348

## Distribuição comprimento do texto



## Word Cloud

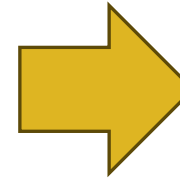


## Métrica de avaliação

$$Accuracy = \frac{Previsões\ Corretas}{Previsões\ Totais}$$

# BASELINE DE FERRAMENTAS EXISTENTES

Ferramenta	<i>Accuracy</i>
TextBlob	70,2%
Pattern	70,2%
SpaCy	70,2%
Vader Sentiment	70,3%
Afinn	71,5%
Flair	90,3%
Text2Emotion	56,2%



## Flair

Destaca-se com 90,3% de accuracy, beneficiando do seu treino específico em reviews do IMDb

- *Pattern e spaCy utilizam o TextBlob internamente, o que justifica a obtenção da mesma accuracy*

# APLICAÇÃO DE LÉXICO DE SENTIMENTOS

**Pré-  
processa  
mento**

Case folding e remoção de pontuação

**NCR Lexicon**

Com negação: 63,3%  
Sem negação: 64,8%

**SentimentWordNet\***

Com negação: 61,6%  
Sem negação: 63,2%

*\*Necessário aplicar POS Tagging*

**Bing Liu Opinion**

Com negação: 72,9%  
Sem negação: 72,5%

# MODELO APRENDIZAGEM AUTOMÁTICA

## Pré-processamento

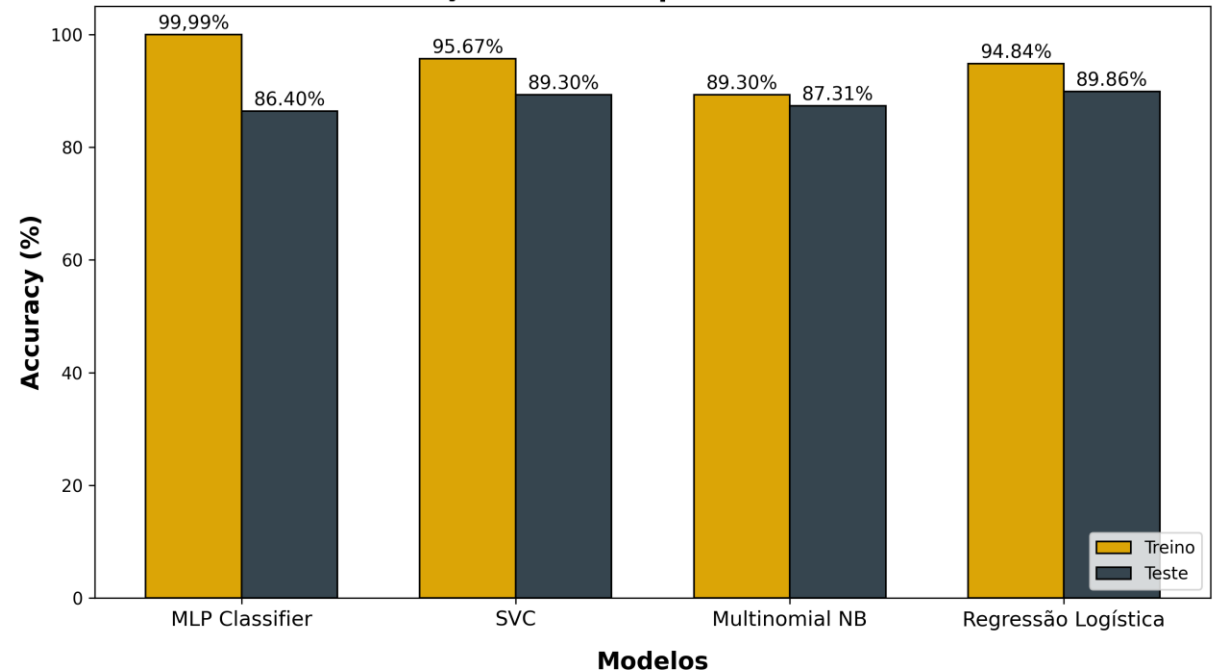
Case folding, remoção de pontuação HTML e URLs, remoção de stopwords, lematização

## Feature Extraction

Método	Accuracy Teste	Accuracy Treino
TF-IDF	0.898618	0.948377
Bow	0.872750	0.998253
Word2Vec	0.807510	0.810656
FastText	0.790144	0.793739

Os métodos foram testados com Regressão Logística num primeiro momento

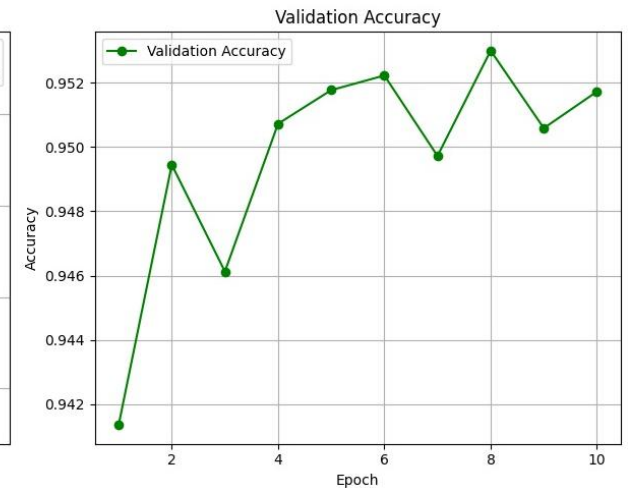
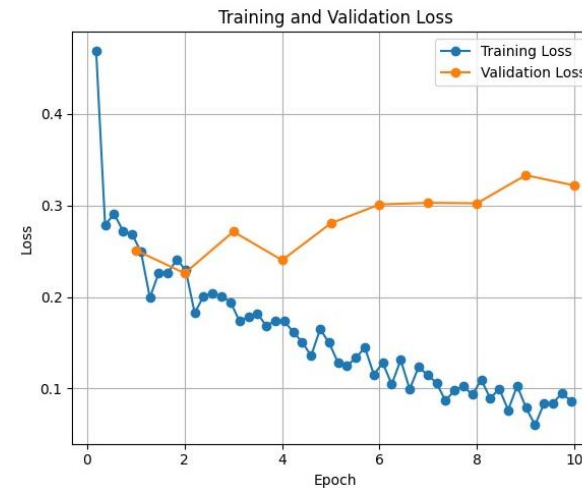
Avaliação de Desempenho dos Modelos



# MODELO TRANSFORMADORES

- O pré-processamento limitou-se à remoção de duplicados e de caracteres especiais (HTML e URLs)

Ferramenta	Accuracy
<i>Sem finetuning</i>	
DistilBERT	90%
TextAttack BERT base uncased IMDb	93,8%
<i>Com finetuning</i>	
Distilbert base	93,5%
RoBERTa base	95,3%

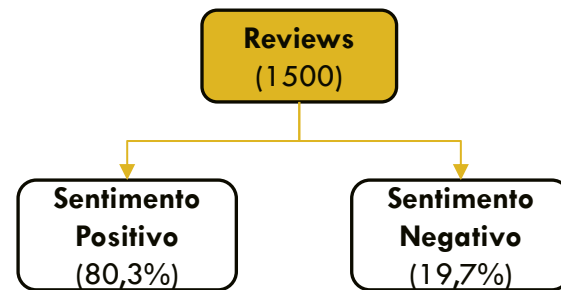


- Todas as experiências observaram overfitting após as primeiras épocas e inconsistências durante o treino



# MODELOS GENERATIVOS

Reviews em PT sobre computadores portáteis



- Usado para inferência, não necessita balanceamento

Ferramenta	Idioma do <i>prompt</i>	Accuracy
Com melhorias no <i>prompt</i>		
Gervásio 7B PT-BR	PT-BR	96,4%
Glória 1.3B	PT-PT	56,7%
Prompts simples		
Granite-3.2-8B-Instruct	EN	93,3%
Mistral-7B-Instruct-v0.1	EN	91,5%

## Prompt simples

```
prompts = [  
    f"""  
        Your task is to analyze the sentiment of this sentence in Portuguese.  
        Respond with only one word: "positive" or "negative".  
        Do not explain. Do not write anything else. Sentence: {texto}  
        Answer: """  
    for texto in batch  
]
```

## Prompt com melhorias

```
prompts = [  
    f"""Classifica o sentimento da seguinte review de um utilizador sobre o desempenho de um computador.  
    INSTRUÇÕES:  
    - Responde apenas com UMA palavra: "positivo" ou "negativo".  
    - Não escrevas mais nada. Não expliques.  
    - Não uses pontuação, espaços ou outros símbolos.  
    Exemplo 1:  
    input: "Notebook bom e rápido, tem um HD espaçoso e uma configuração que atende muito bem para o trabalho diário..."  
    Classificação: positivo  
    Exemplo 2:  
    input: "Atende as minhas necessidades. Muito bom. Roda bem os programas, tem um belo design..."  
    Classificação: positivo  
    Exemplo 3:  
    input: "Eu fiz a compra mês passado e cancelei às vésperas de receber o produto. Motivo: fiz a compra em boleto..."  
    Classificação: negativo  
    Exemplo 4:  
    input: "Comprei o produto após ver alguns vídeos na internet sobre seu funcionamento... porém não funciona com IOS11..."  
    Classificação: negativo  
    input: "{texto}"  
    Classificação: ""  
    for texto in batch  
]
```

# CONCLUSÃO E DISCUSSÃO

- A ordem e escolha das técnicas de pré-processamento influenciam o desempenho dos modelos.
- Nem sempre um pré-processamento mais exaustivo resulta em melhores desempenhos, o ajuste ao contexto é essencial.
- Modelos mais simples podem ter resultados interessantes, como é o caso da Regressão Logística com TF-IDF.
- Transformadores, como o RoBERTa, têm melhores resultados, mas exigem mais recursos computacionais (nomeadamente, subscrição do Google Colab ou placa gráfica Nvidia).
- Para aplicações que exigem rapidez, modelos clássicos continuam a ser a escolha mais viável.
- A escolha do prompt tem um alto impacto nos modelos generativos, além da língua e o números de tokens.

# PROPOSTAS PARA TRABALHO FUTURO

- Explorar redes neurais convolucionais (CNN) e recorrentes (RNN);
- Avaliar impacto de técnicas adicionais (e.g., correção ortográfica, detecção de ironia, análise de contexto);
- Aplicar técnicas como grid search e random search para o ajuste dos modelos baseados em transformadores;
- Testar diferentes hiperparâmetros (weight decay, dropout, early stopping) + teste de outros modelos nos transformadores;
- Mais experiências para os modelos generativos (variações e idioma dos prompts, estratégias como prompt-chaining, número de tokens, etc.);
- Desenvolver métodos específicos para mitigar os casos onde os modelos falham, e.g., textos ambíguos, críticas mistas, e ironia.

# ANEXOS

## Modelo Aprendizagem Automática

Textos classificados incorretamente:

- Falsos Positivos

Texto: *“Disappointing film with Walter Pidgeon [...] The story gets bogged down somewhat as love develops between the two [...] The near ending scene in the cave is nicely realized [...] many co stars that same year in the memorable How Green Was My Valley.”*

Sentimento real: **negativo**

Sentimento predito: **positivo**

- Falsos Negativos

Texto: *“Witchery is a decent little Euro Trash horror yarn! David Hasselhoff is pretty damn funny in this one and sadly, he's one of the better actors. Linda Blair is downright terrible and the lady who plays Hoff's wife...she is hilariously bad! The plot of this film is ridiculous too. It's got some holes [...] but it remains entertaining throughout. The gore in "Witchery" is freaking outstanding. I loved the part where the old bag gets "sucked" into the trash chute [...] I'd say this one is a must for gorehounds. [...] 7 out of 10, kids.”*

Sentimento real: **positivo**

Sentimento predito: **negativo**

# ANEXOS

## Modelos Generativos

Identificados dois tipos de respostas incorretas (modelo Gervásio):

- Classificações incorretas

Texto: “*Prós: Marca renomada/possível boa durabilidade. Contras: processador lento/ aparelho pesado/ bateria dura pouco/ não vem com Office*”

Sentimento real: **negativo**

Sentimento predito: **positivo**

- Respostas incoerentes

Texto: “*Alguém sabe se é bom para rodar o Premiere?? Estou procurando um notebook para trabalho.*”

Sentimento real: **positivo**

Sentimento predito: “**não específico**”