Fato ou Fake?

Combatendo desinformação com Aprendizagem de Máquina e Processamento de Linguagem Natural

@Thais Almeida
Data Scientist

Sobre mim

Mestre em Computação pela Universidade Federal do Amazonas (UFAM). Tem 4 anos de experiência trabalhando com **Social Network Analysis** e **Natural Language Processing**. Entre suas contribuições acadêmicas, desenvolveu pesquisas relacionadas com a investigação de **depoimentos legais** (Operação Lava Jato) e detecção de **discursos de ódio** e **notícias falsas** em ambientes online. Atualmente, trabalha como Cientista de Dados na **Creditas**.

O que são “fake news”?

creditas

Definições

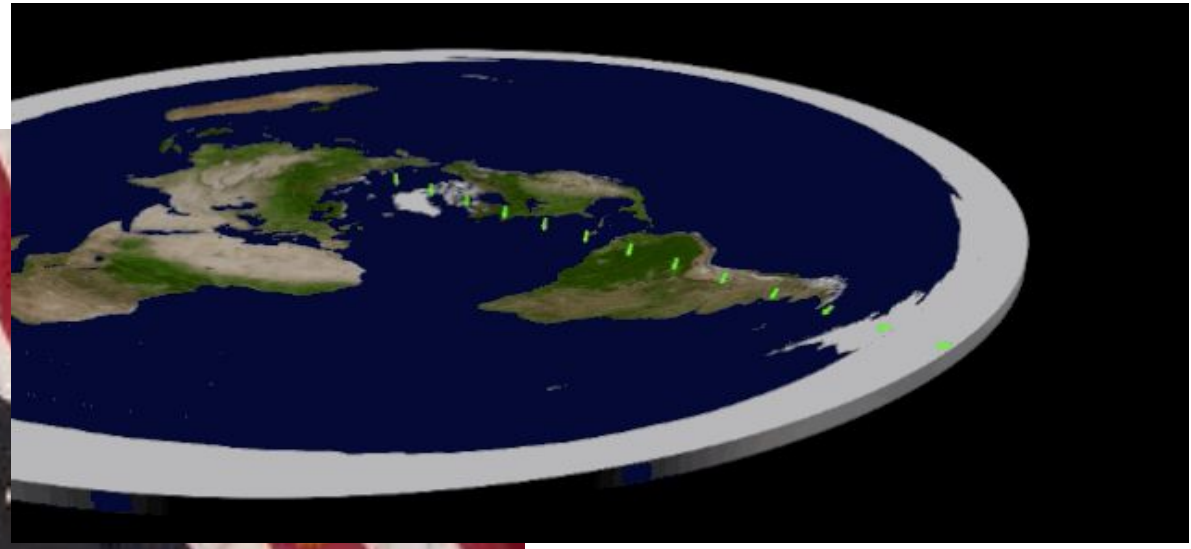
Fake news são informações fabricadas que imitam o conteúdo da mídia na forma, mas não no processo organizacional ou intenção.

(Science, 2018)

Notícias falsas são artigos de notícias cujo conteúdo é intencionalmente e verificadamente falso.

(Allcott e Gentzknow, 2017)

Por que é necessário combater “fake news”?



Do Photos Show Greta Thunberg with George Soros, ISIS, and the 'Antifa Terrorist Organization'?

24 SEPTEMBER 2019 | FACT CHECK

The teenage climate activist's high-profile speeches in Congress an the United Nations prompted attacks from online detractors.

Candidato a embaixador, Eduardo Bolsonaro publica foto falsa de Greta

Brasil 247

5 Top Fake Political News Stories On Facebook

"Obama Signs Executive Order Banning The Pledge Of Allegiance In Schools Nationwide" ABCNews.com.co

2,177,000 Facebook shares, comments, and reactions

"Pope Francis Shocks World, Endorses Donald Trump for President, Releases Statement" Ending the Fed

961,000

"Trump Offering Free One-Way Tickets to Africa & Mexico for Those Who Wanna Leave America" tmzhiphop.com

802,000

"FBI Agent Suspected in Hillary Email Leaks Found Dead in Apparent Murder-Suicide" Denver Guardian

567,000

"RAGE AGAINST THE MACHINE To Reunite And Release Anti Donald Trump Album" heaviermetal.net

560,000

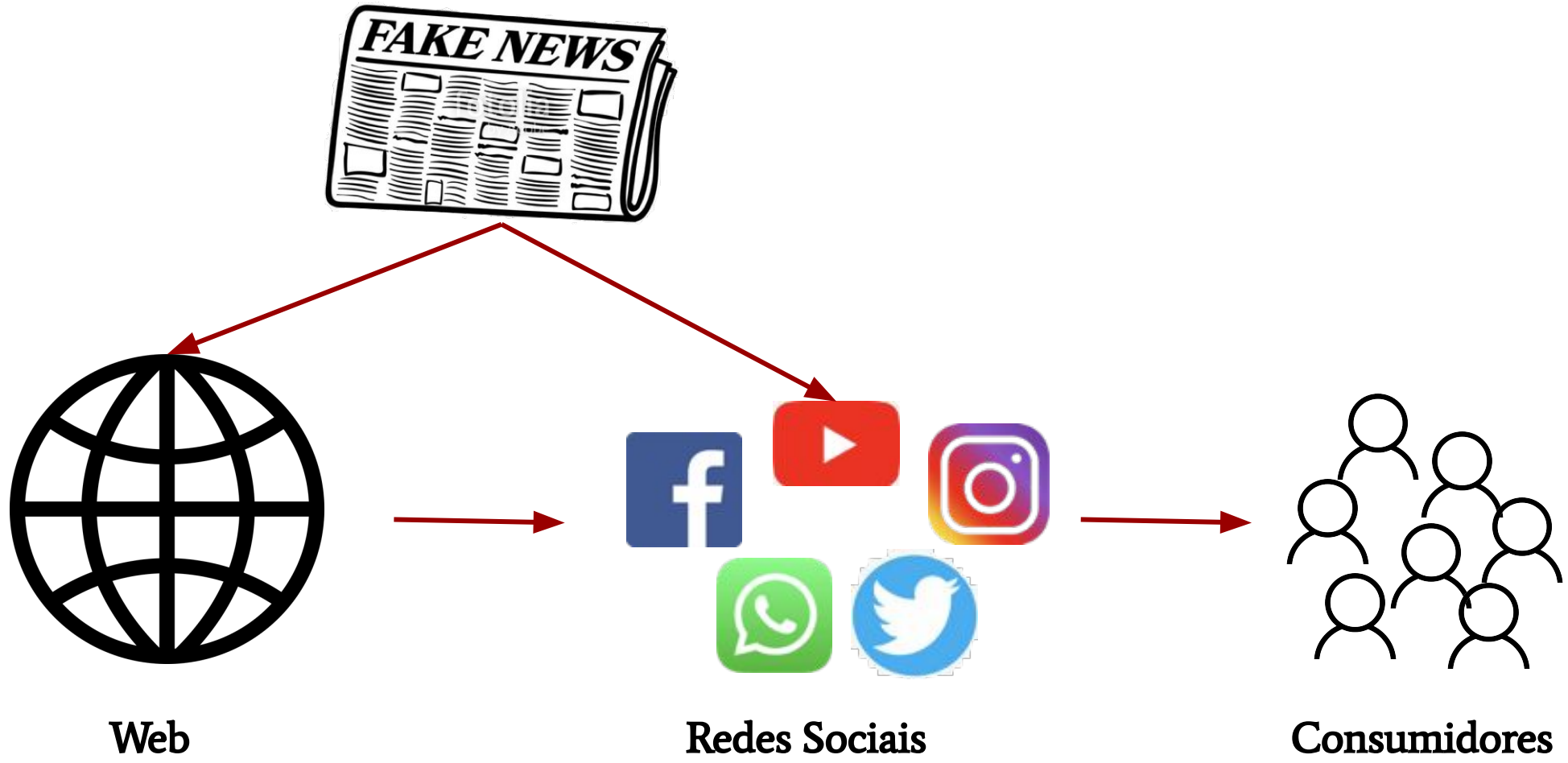
Brasil

Desembargadora diz que Marielle 'estava engajada com bandidos'

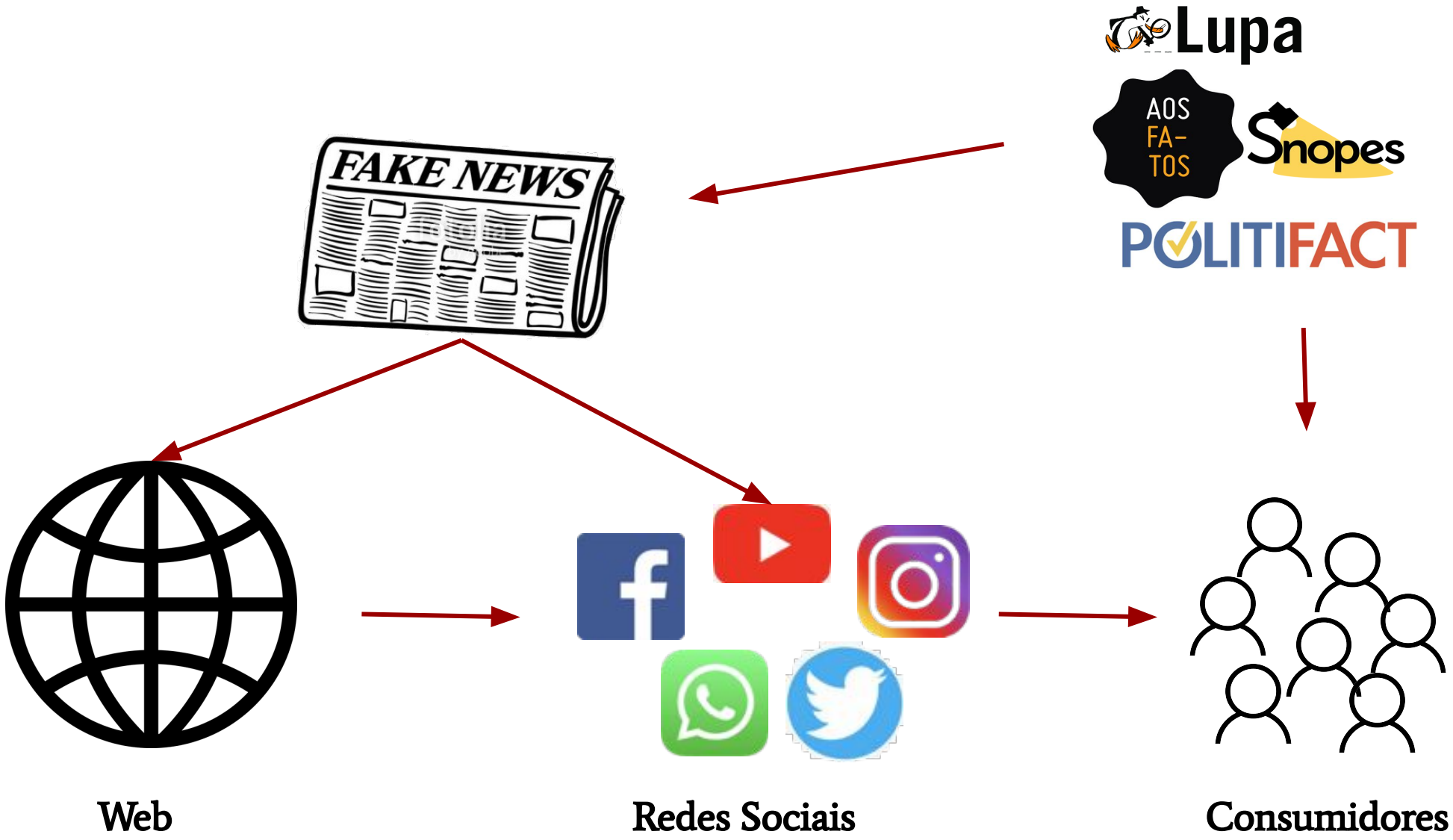
Com base em fake news, Marília Castro Neves, do TJ-RJ, afirmou em comentário que vereadora morta 'foi eleita pelo Comando Vermelho'

creditas

Visão Geral: Ecossistema de Desinformação



Visão Geral: Ecossistema de Desinformação



Visão Geral: Ecossistema de Desinformação

True	Mostly True
Mixture	Mostly False
False	Unproven
Outdated	Miscaptioned
Correct Attribution	Misattributed
Scam	Legend

Visão Geral: Ecossistema de Desinformação (resuminho)

- Principais motivadores: **financeiro** e **político**.
- Os desafios ao lidar com desinformação são **qualitativos** e **quantitativos**.
- **Métodos automáticos** podem auxiliar na **verificação de fatos**, fornecendo **informações contextuais** e **limitando o volume de dados**.

Qual estratégia utilizar para detectar fake news?

creditas

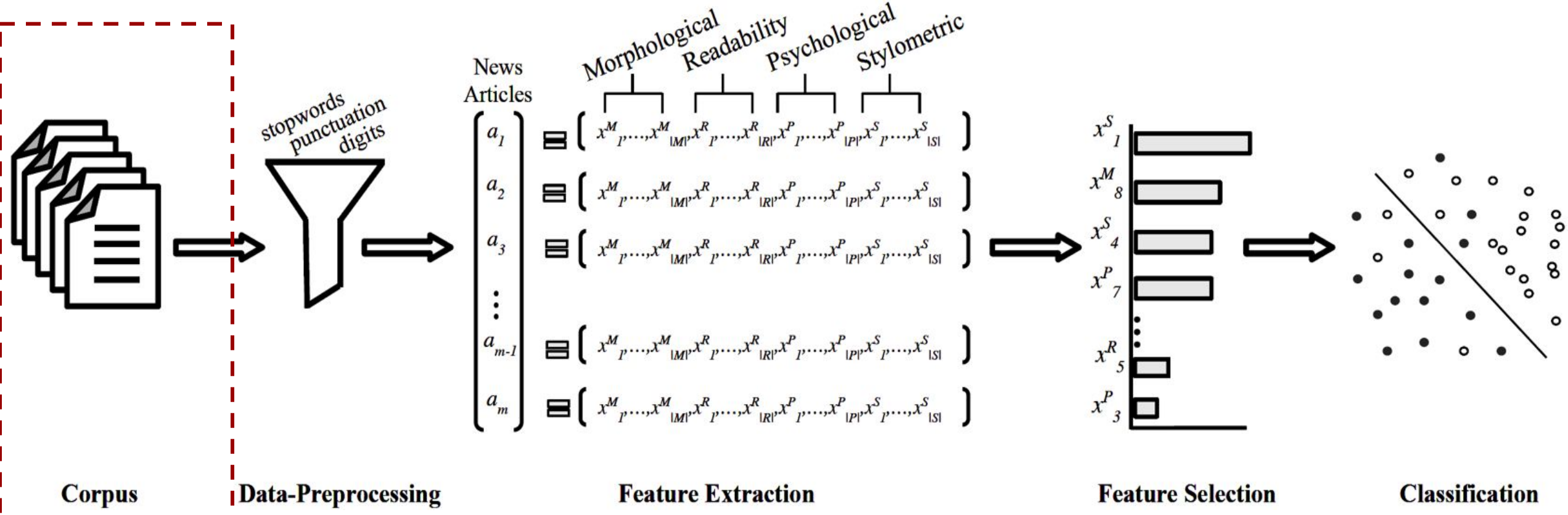
Deixando um pouco mais formal:

Dado um **artigo de notícia A** , o objetivo é determinar **automaticamente** se **A é falso ou legítimo**.

Falso

Legítimo

Visão geral: LiarDetector



creditas

Desafios com bases de dados

Dataset	Source	Ground truth	#samples	Text metadata	Social metadata	Publishers metadata
BS Detector	Web	Chrome Plugin	12,999	✓	✓	
BuzzFeed-Webis	Web, Facebook	Journalists	1,627	✓	✓	✓
Celebrity	Web	Gossip checker	500	✓		
Credbank	Twitter	Crowdsourcing	1049	✓	✓	
Emergent	Web	Journalists	1,600	✓		
Liar	PolitiFact	PolitiFact	12,836	✓		
NewsReliability	Web	Journalists	74,476	✓		
FakeNewsNet	Web, Twitter	BuzzFeed, Politifact	422	✓	✓	✓
US-Election2016	Web, Facebook	Snopes, Politifact	948		✓	
Fake.br	Web	Authors	7,000	✓	✓	✓

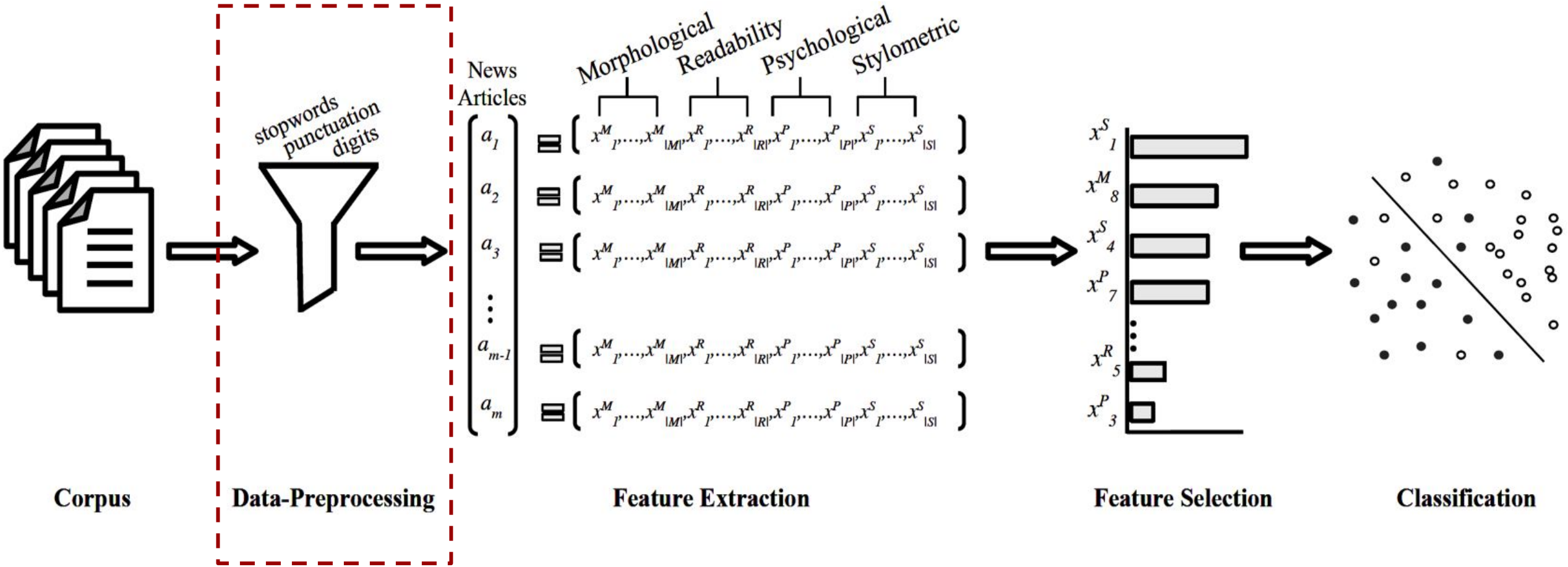
Desafios com bases de dados

Dataset	Source	Ground truth	#samples	Text metadata	Social metadata	Publishers metadata
BS Detector	Web	Chrome Plugin	12,999	✓	✓	
BuzzFeed-Webis	Web, Facebook	Journalists	1,627	✓	✓	✓
Celebrity	Web	Gossip checker	500	✓		
Credbank	Twitter	Crowdsourcing	1049	✓	✓	
Emergent	Web	Journalists	1,600	✓		
Liar	PolitiFact	PolitiFact	12,836	✓		
NewsReliability	Web	Journalists	74,476	✓		
FakeNewsNet	Web, Twitter	BuzzFeed, Politifact	422	✓	✓	✓
US-Election2016	Web, Facebook	Snopes, Politifact	948		✓	
Fake.br	Web	Authors	7,000	✓	✓	✓

Desafios com bases de dados

Dataset	Source	Ground truth	#samples	Text metadata	Social metadata	Publishers metadata
BS Detector	Web	Chrome Plugin	12,999	✓	✓	
BuzzFeed-Webis	Web, Facebook	Journalists	1,627	✓	✓	✓
Celebrity	Web	Gossip checker	500	✓		
Credbank	Twitter	Crowdsourcing	1049	✓	✓	
Emergent	Web	Journalists	1,600	✓		
Liar	PolitiFact	PolitiFact	12,836	✓		
NewsReliability	Web	Journalists	74,476	✓		
FakeNewsNet	Web, Twitter	BuzzFeed, Politifact	422	✓	✓	✓
US-Election2016	Web, Facebook	Snopes, Politifact	948		✓	
Fake.br	Web	Authors	7,000	✓	✓	✓

Visão geral: LiarDetector



creditas

Pré-processamento de Dados

Título (headline)

DISGUSTING! Because Of Hillary & Obama, NY Terrorist Will Get Better Treatment Than US Vets! - Freedom Daily

What a sign of the times it is when the terrorist who planted and detonated a number of explosives in New Jersey and New York City Saturday will be receiving better care than our own veterans who've fought to protect ourselves from men like him. It sounds unreal, but that's the truth in Obama's America.

Corpo (content)

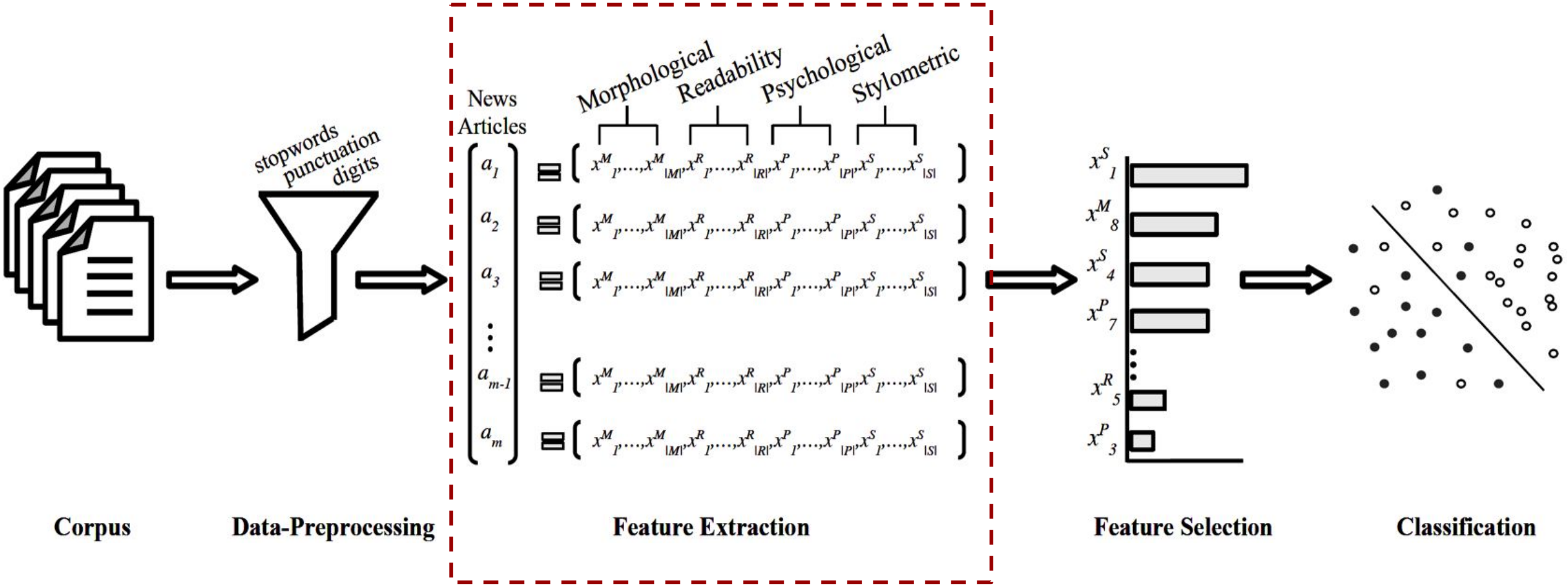
creditas

Pré-processamento de Dados

Entrada: DISGUSTING! Because Of Hillary & Obama, NY Terrorist Will Get Better Treatment Than US Vets! - Freedom Daily

Saída: [disgusting, hillary, obama, ny, terrorist, get, better, treatment, us, vets]

Visão geral: LiarDetector



creditas

	Description	Description	Description
Morphological Features	Conjunction, coordinating	Pre-determiner	Interjection
	Numeral, cardinal	Verb, past tense	Verb, base form
	Determiner	Noun, proper, plural	Verb, present participle or gerund
	Foreign word	Noun, common, plural	Verb, past participle
	Preposition or conjunction, subordinating	Genitive marker	Verb, present tense, not 3rd singular
	Adjective or numeral, ordinal	Pronoun, personal	Verb, present tense, 3rd singular
	Adjective, comparative	Pronoun, possessive	WH-determiner
	Adjective, superlative	Adverb	WH-pronoun
	Modal auxiliary	Adverb, comparative	WH-pronoun, possessive
	Noun, common, singular or mass	Adverb superlative	Wh-adverb
	Noun, proper, singular	"To" as preposition/infinitive	Particle
Psychological Features	Summary Dimensions (word tone)	Biological Processes (ingest, health)	Affect (anger, sad, anxiety)
	Function Words (pronoun, negations)	Drives (power, risk)	Relativity (space, time)
	Punctuation Marks (comma, semicolon)	Other Gramar (quantifiers, interrogatives)	Personal Concerns (home, work)
	Perceptual Process (see, hear)	Time Orientation (focuspast, focuspresent)	Social (family, friend)
	Cognitive Processes (insight, certainty)	Informal Language (netspeak, filler)	
Readability Features	Flesch Reading Ease	Words per sentence	Long words
	Flesch Kincaid Grade	Capitalized words	Syllables
	McLaughlin's SMOG	Percentage of stopwords	Lexicon
	Gunning Fog	Urls	Sentences
	Coleman-Liau	Difficult words	Words
	Automated Readability	Characters	
	Linsear Write	Complex words	
Stylometric features	Jensen Shannon divergence	Normalized Shannon entropy	

	Description	Description	Description
Morphological Features	Conjunction, coordinating Numeral, cardinal Determiner Foreign word Preposition or conjunction, subo Adjective or numeral, ordinal Adjective, comparative Adjective, superlative Modal auxiliary Noun, common, singular or mass Noun, proper, singular	Pre-determiner Verb, past tense Noun, proper, plural ural Pronoun, possessive Adverb Adverb, comparative Adverb superlative "To" as preposition/infinitive	Interjection Verb, base form Verb, present participle o Verb, past participle Verb, present tense, not 3rd sin Verb, present tense, 3rd singula WH-determiner WH-pronoun WH-pronoun, possessive Wh-adverb Particle
Psychological Features	Summary Dimensions (word tone) Function Words (pronoun, negations) Punctuation Marks (comma, semicolon) Perceptual Process (see, hear) Cognitive Processes (insight, certainty)	Biological Processes (ingest. health) Drives (power, risk) Other Gramar (quantifiers, Time Orientation (focuspas Informal Language (netspeak, filler)	Affect (anger, sad, anxiety) ace, time) cerns (home, work) ; friend)
Readability Features	Flesch Reading Ease Flesch McLau Gunni Coleman-Liau Automated Readability Linsear Write	Words per sentence Capitalized words Percentage of stopwords Urls Difficult words Characters Complex words	Long words Syllables Lexicon Sentences Words
Stylometric features	Jensen Shannon divergence	Normalized Shannon entropy	

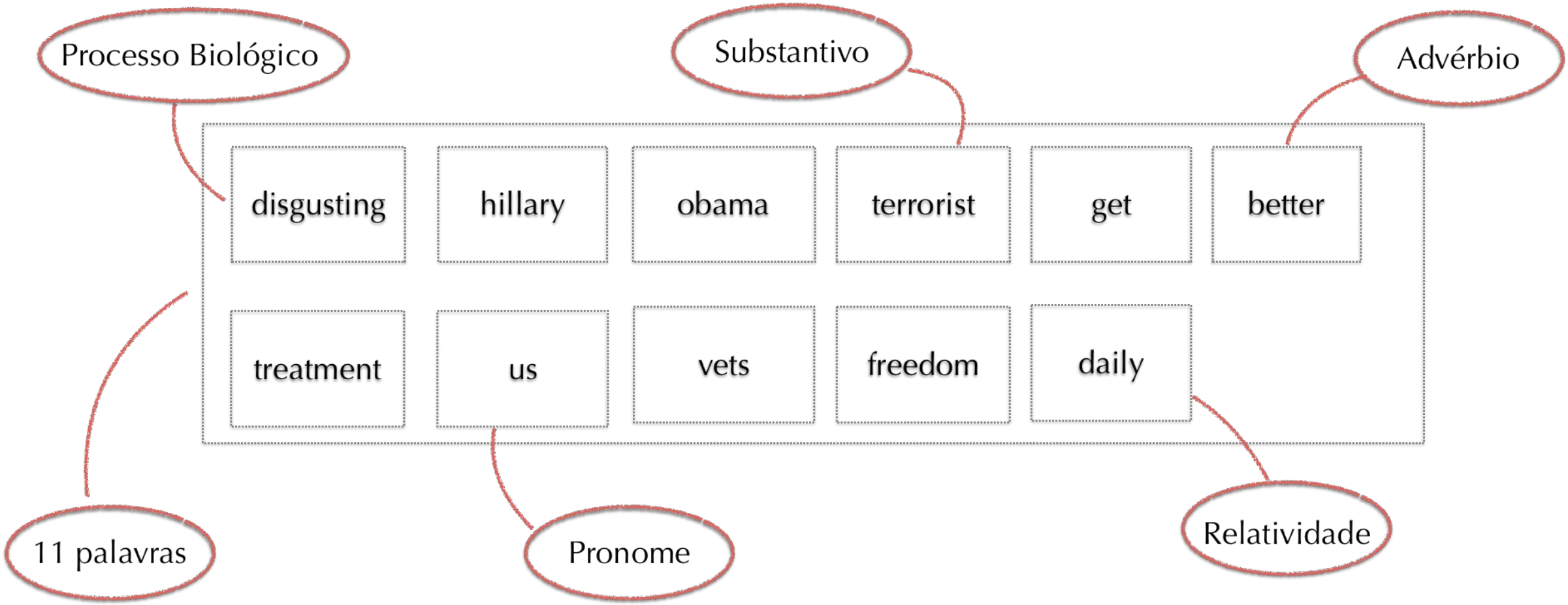
NLTK

LIWC

Textstat

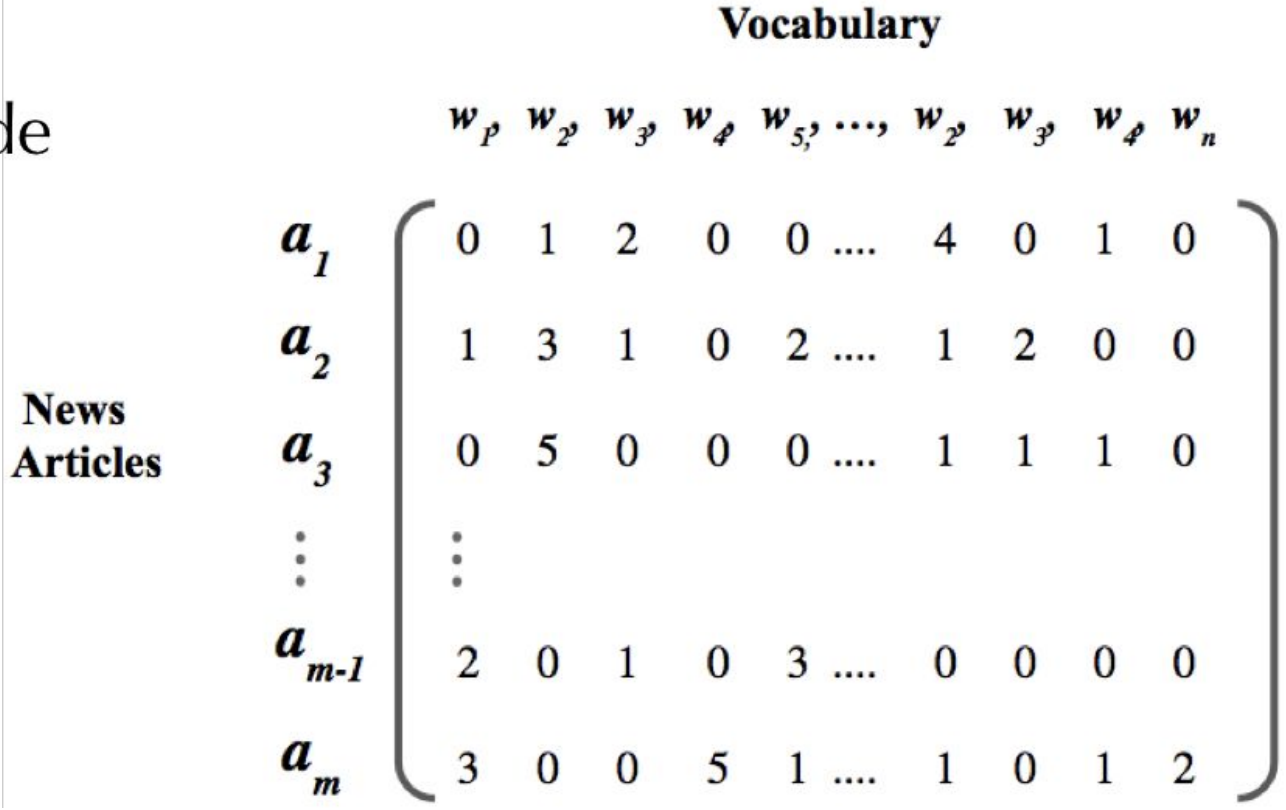
Scipy

Extração de Features: morfológicas, psicológicas, legibilidade



Extração de Features: estilométricas (motivação)

Trabalhos anteriores representaram notícias principalmente por meio de **n-grams**.



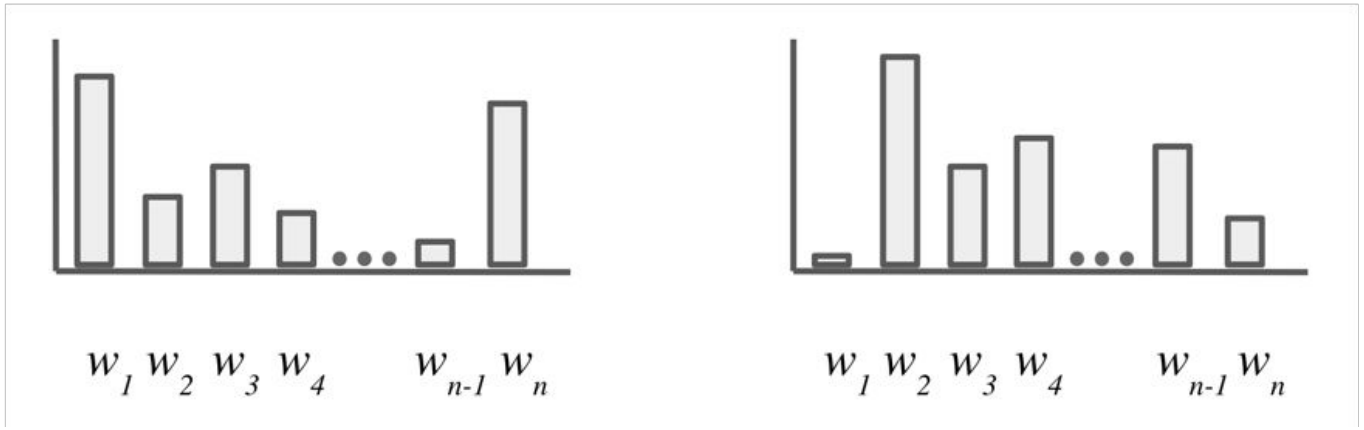
creditas

Extração de Features: estilométricas

Artigos

	w_1	w_2	w_3	w_4	w_5	...	w_2	w_3	w_4	w_n
a_1	0	1	2	0	0	....	4	0	1	0
a_2	1	3	1	0	2	....	1	2	0	0
a_3	0	5	0	0	0	....	1	1	1	0
⋮	⋮									
a_{m-1}	2	0	1	0	3	....	0	0	0	0
a_m	3	0	0	5	1	....	1	0	1	2

Matriz de Termos

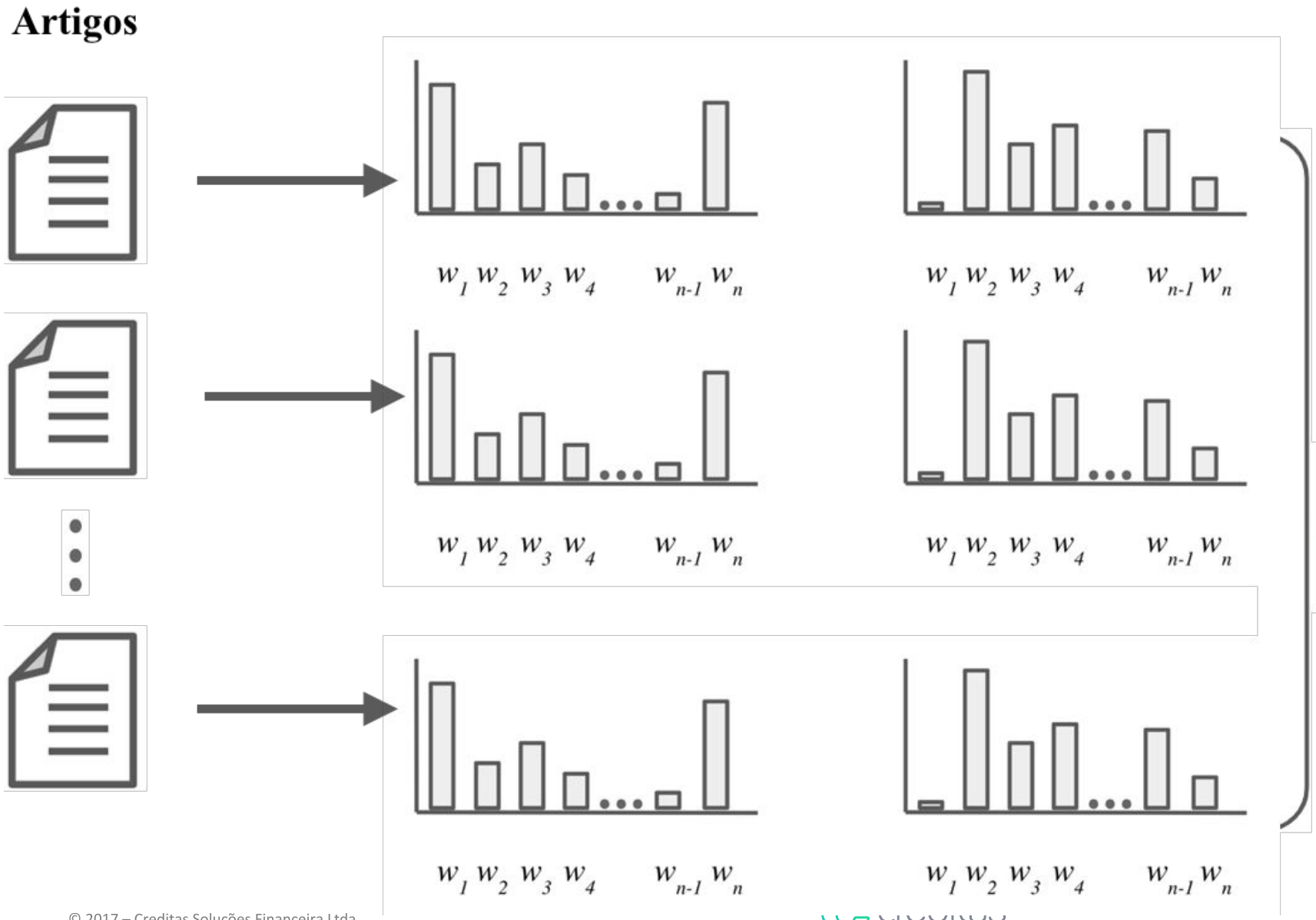


Histogramas de Referência

$$\left\langle p_i^{(c)} \right\rangle = \left\langle f_i^{(c)} \right\rangle \Big/ \sum_{i=1}^{N} \left\langle f_i^{(c)} \right\rangle$$

creditas

Extração de Features: estilométricas



Extração de Features: estilométricas

Shakespeare***

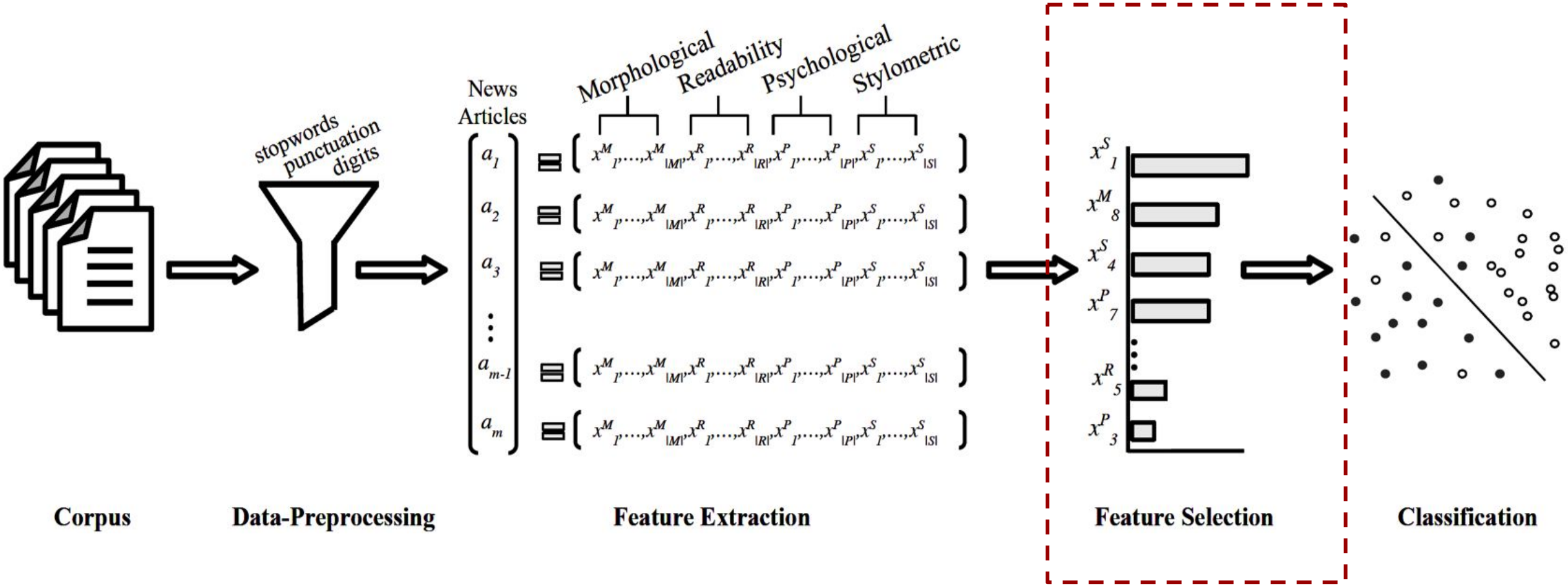
Entropia de Shannon Normalizada

$$H(P) = S(P)/S_{max} = \left(-\sum_{p_i \epsilon P} p_i \log(p_i)\right) \Big/ S_{max}$$

Divergência de Jensen-Shannon

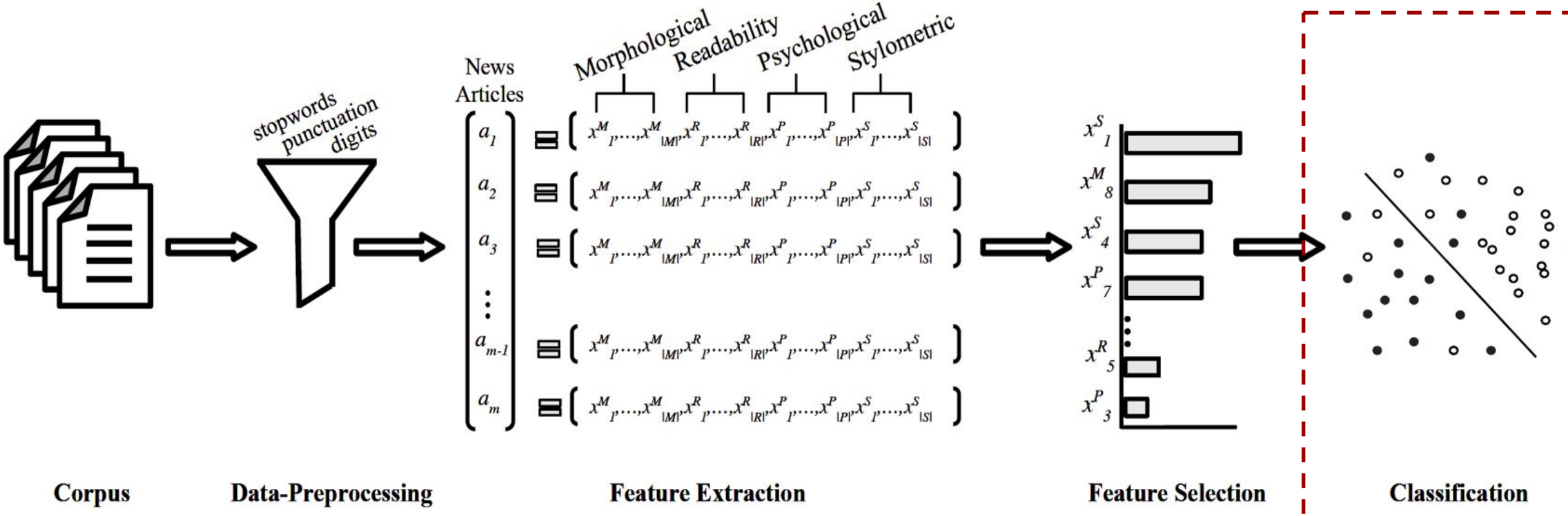
$$JSD(P,Q) = S\left(\frac{P+Q}{2}\right) - \frac{S(P)+S(Q)}{2}$$

Visão geral: LiarDetector



creditas

Visão geral: LiarDetector



creditas

Avaliação Experimental

Bases de Dados:

- **Celebrity**, FakeNewsNet
- Emergent, Fake.br

Baseline:

- **FNDetector** (Pérez et al., 2018);

Algoritmos de Aprendizagem:

- Support Vector Machine (**SVM**)
- K-Nearest Neighbor (**KNN**)
- Random Forest (**RNF**)
- Gaussian Naive Bayes (**GNB**)

Avaliação Experimental

Métricas de Avaliação

- **Precisão e Revocação**
- **F1**

Bibliotecas:

- **LIWC, NLTK, TextStat, GoogleTrans, Standford Parser, Scikit-learn**

Avaliação dos Modelos:

- **10-fold cross validation**
- **Leave-one-out**

Configurações de Hardware:

- **MacOS, Intel Core I7 2.5GHz, 16GB.**

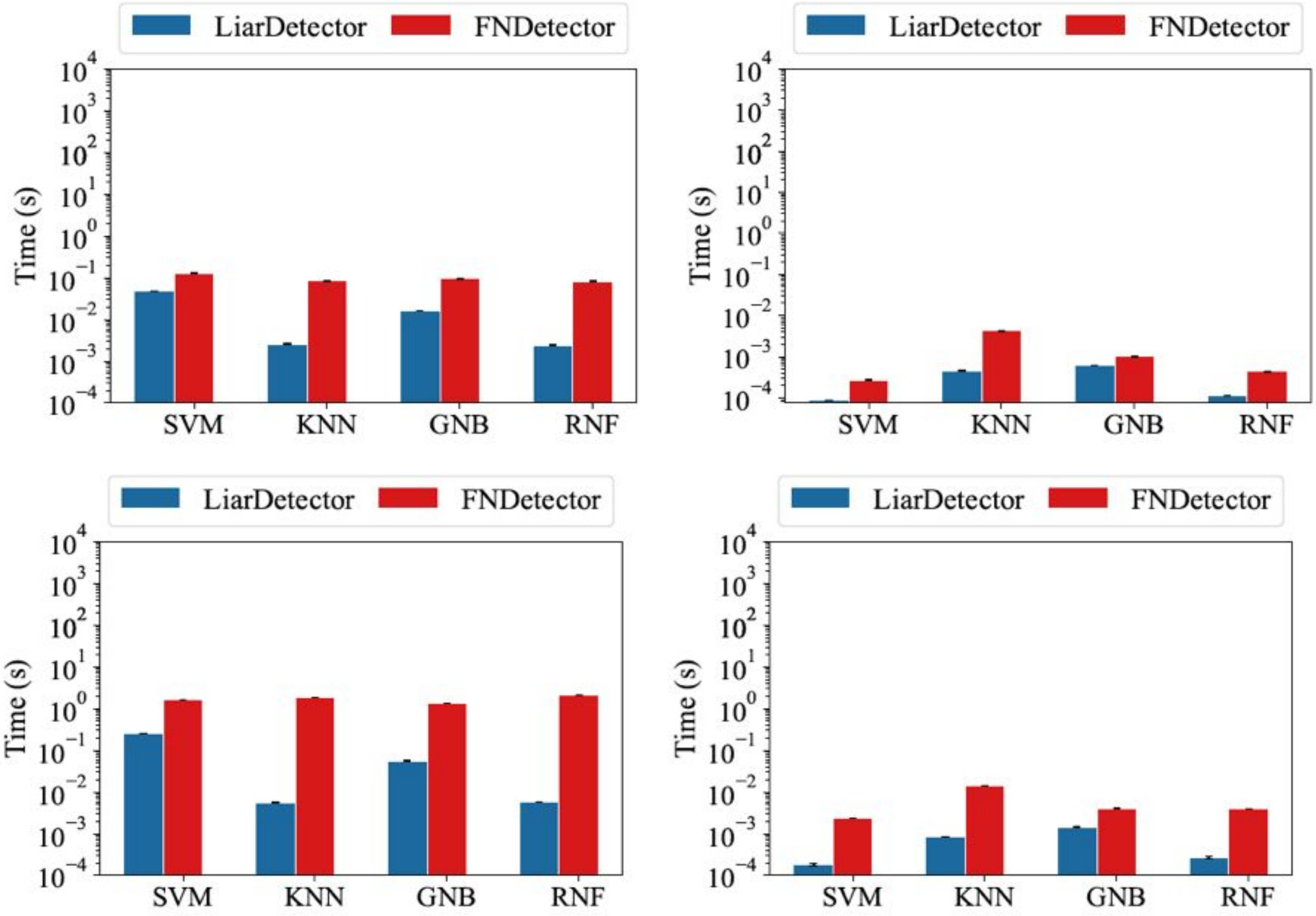
Avaliação Experimental: Celebrity (dataset)

Text	Approach	Metric	Support Vector Machine		K-Nearest Neighbor		Gaussian Naive Bayes		Random Forest	
			Fake	Real	Fake	Real	Fake	Real	Fake	Real
Headline	H+JSD	PR	0.62	0.61	0.61	0.6	0.61	0.67	0.54	0.55
		RE	0.61	0.62	0.58	0.63	0.75	0.52	0.67	0.41
		F1	0.62	0.61	**0.60**	**0.61**	**0.68**	**0.59**	0.6	0.47
	FNDetector	PR	0.64	0.64	0.61	0.6	0.38	0.35	0.61	0.63
		RE	0.65	0.62	0.58	0.63	0.4	0.33	0.67	0.55
		F1	**0.64**	**0.63**	**0.60**	**0.61**	0.39	0.34	**0.64**	**0.59**
Content	H+JSD	PR	0.64	0.65	0.64	0.68	0.68	0.79	0.59	0.58
		RE	0.65	0.64	0.71	0.61	0.84	0.61	0.56	0.61
		F1	**0.65**	**0.64**	**0.68**	**0.64**	**0.75**	**0.69**	0.57	0.59
	FNDetector	PR	0.64	0.65	0.61	0.65	0.52	0.71	0.66	0.74
		RE	0.67	0.62	0.7	0.56	0.94	0.14	0.79	0.59
		F1	**0.65**	**0.64**	0.65	0.61	0.67	0.23	**0.72**	**0.65**

Treino | Teste

Headline

Content

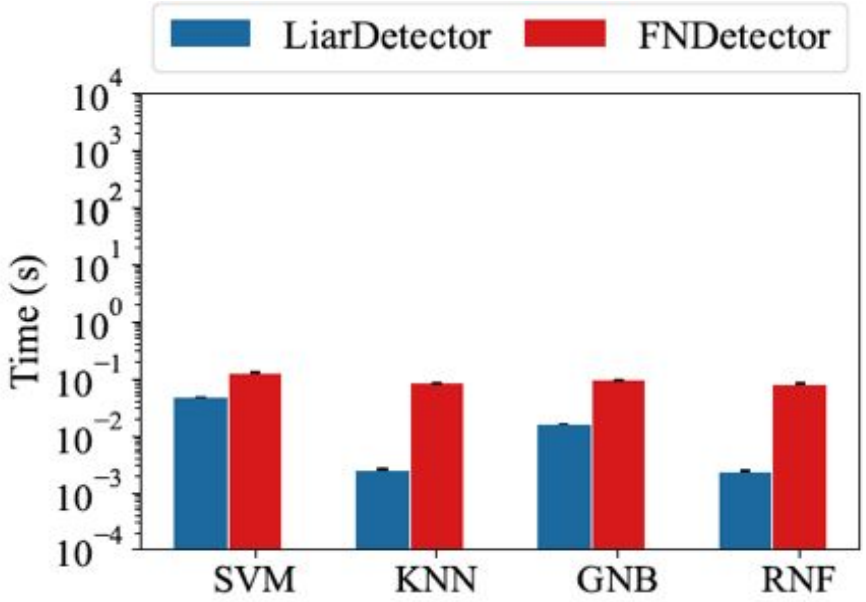


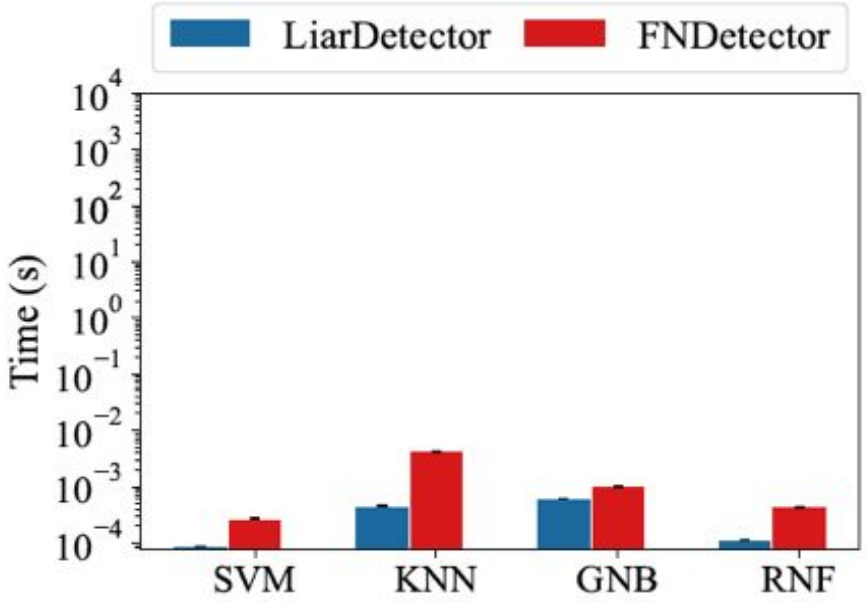
creditas

Treino

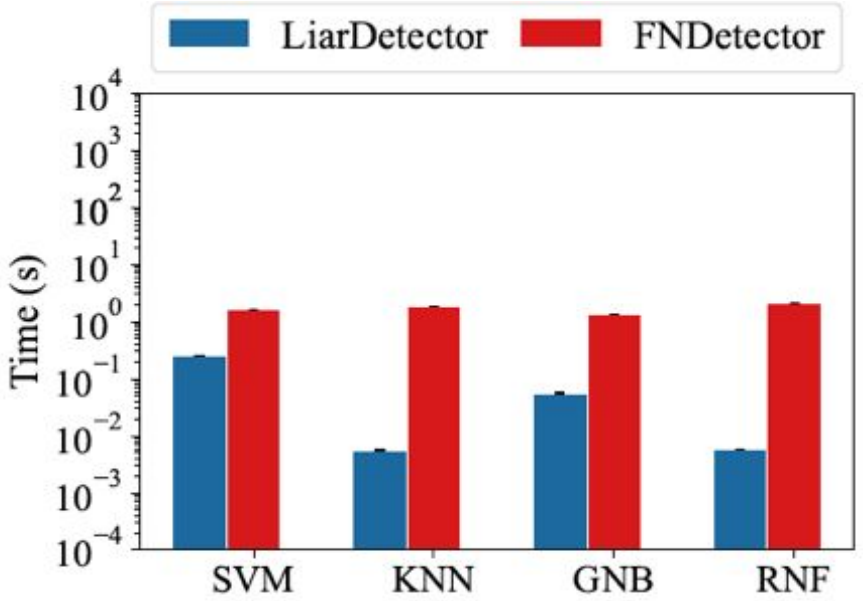
Teste

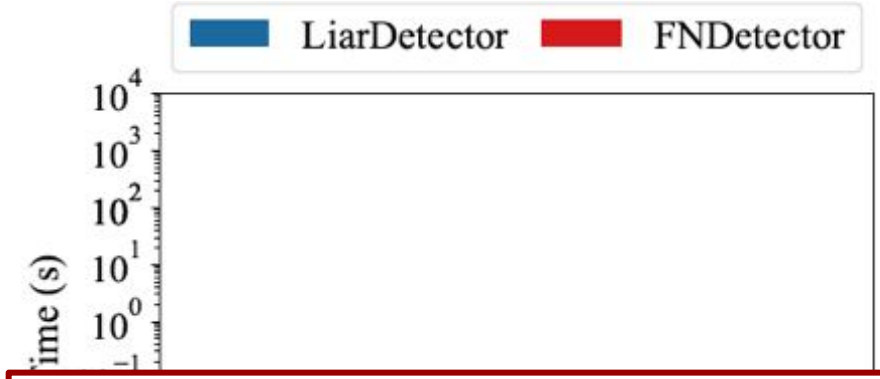
Headline





Content





LiarDetector é até 354 vezes mais rápido que o baseline.

6000**

creditas

É isso ai galera, ML + NLP **ajudam** a identificar *fake news* !!!

....Mas quais seriam as **limitações** desse tipo de abordagem? 🤔

creditas

E o que você pode fazer?

Obrigada =D

Dúvidas?
Sugestões ?
Elogios ?

creditas