XXV ENCONTRO DE
INICIAÇÃO CIENTÍFICA

X ENCONTRO DE INICIAÇÃO EM
DESENVOLVIMENTO TECNOLÓGICO
E INOVAÇÃO

Identificação de Linguagem Ofensiva em Mídias Sociais Utilizando Algoritmos de Aprendizado de Máquina.

Murilo de Paula Araujo[1] Orientador: Juan Manuel Adán Coello[2]

Modalidade: PIBIC/CNPq
[1]Curso de Engenharia de Computação
[2]Grupo de Pesquisa em Sistemas Inteligentes (GPqSI) - CEATEC
Área/Sub-área de conhecimento: Ciências Exatas e da Terra/Ciência da Computação

PROPESQ
Pró-Reitoria de Pesquisa
e Pós-Graduação

PUC
CAMPINAS
PONTIFÍCIA UNIVERSIDADE CATÓLICA

Introdução

- Casos de agressão e abuso *online* têm criado problemas de diversa natureza e gravidade aos usuários, levado muitos deles à desativação de contas, à autoflagelação e até mesmo ao suicídio [1].

Objetivo

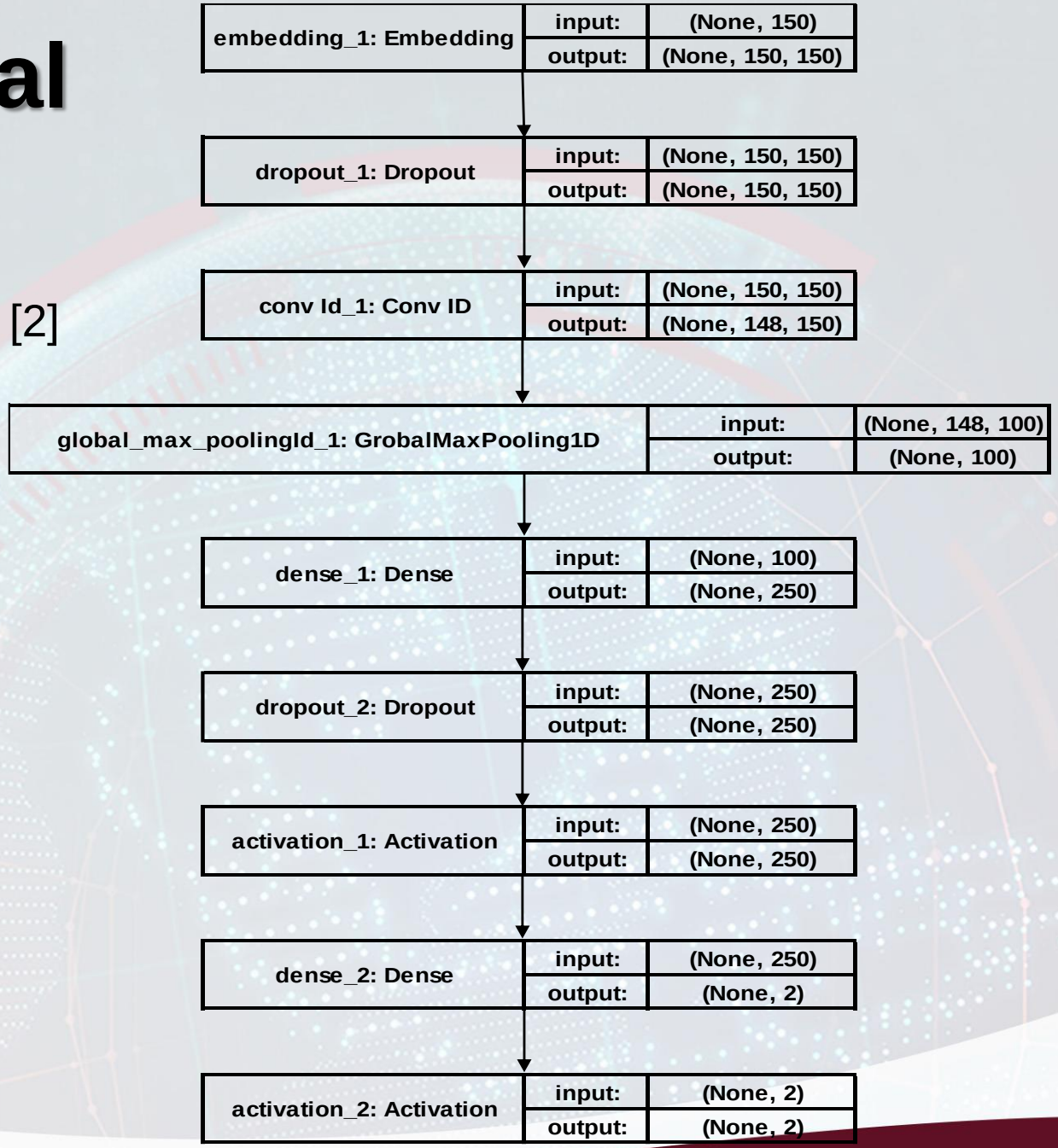
- Produzir modelos de classificação para identificar linguagem ofensiva em textos publicados em mídias sociais, utilizando algoritmos de aprendizado de máquina supervisionado.

Método

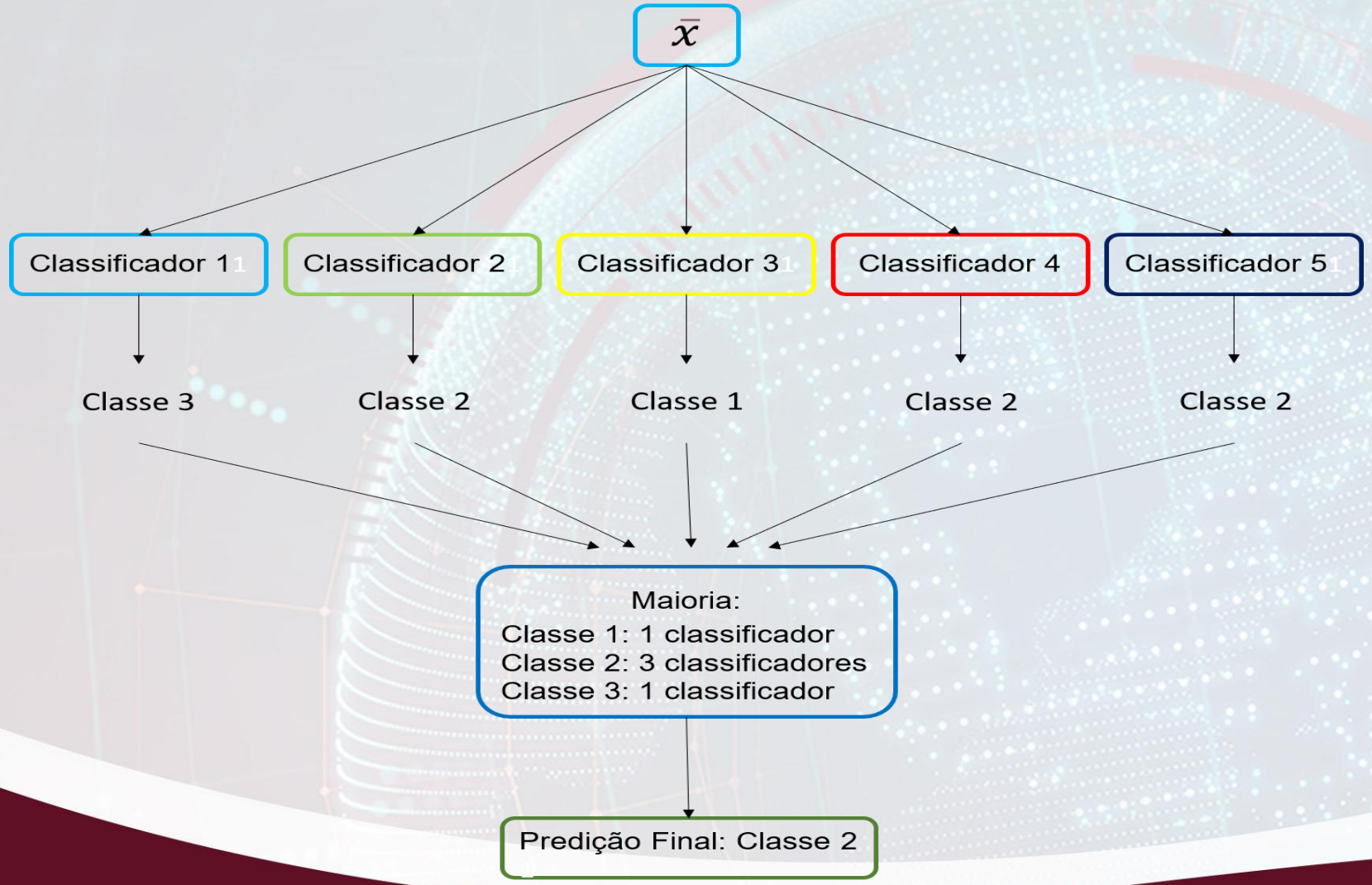
- Obtenção de conjuntos de dados rotulados contendo linguagem ofensiva e não ofensiva originária de mídias sociais.
- Construção de comitês de classificadores para a detecção de linguagem ofensiva em mídias sociais usando redes neurais convolucionais.
- Avaliação dos classificadores utilizando as métricas Acurácia, Medida F, ROC-AUC.

Método – Rede Neural Convolucional

- Desenvolvida em trabalho de IC anterior [2]
- *Camadas:*
  - *Input*
  - *Embedding*
  - *Dropout*
  - *Conv1D*
  - *GlobalMaxPooling1D*
  - *Dense*
  - *Activation*



Método – Comitês de classificadores C-GPqSI [3]



Métricas

		Resposta Esperada	
		Positivo	**Negativo**
Resposta Prevista	**Positivo**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Negativo	Falso Positivo (FP)	Verdadeiro Negativo (VN)

$$\boldsymbol{Acurácia} = \frac{VP + VN}{Total}$$

$$\boldsymbol{Precisão} = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FP}}$$

$$\boldsymbol{Recall} = \frac{VP}{VP + FN}$$

$$\boldsymbol{F1} = \frac{2 * Precisão * Recall}{Precisão + Recall}$$

Resultados – Conjuntos de dados

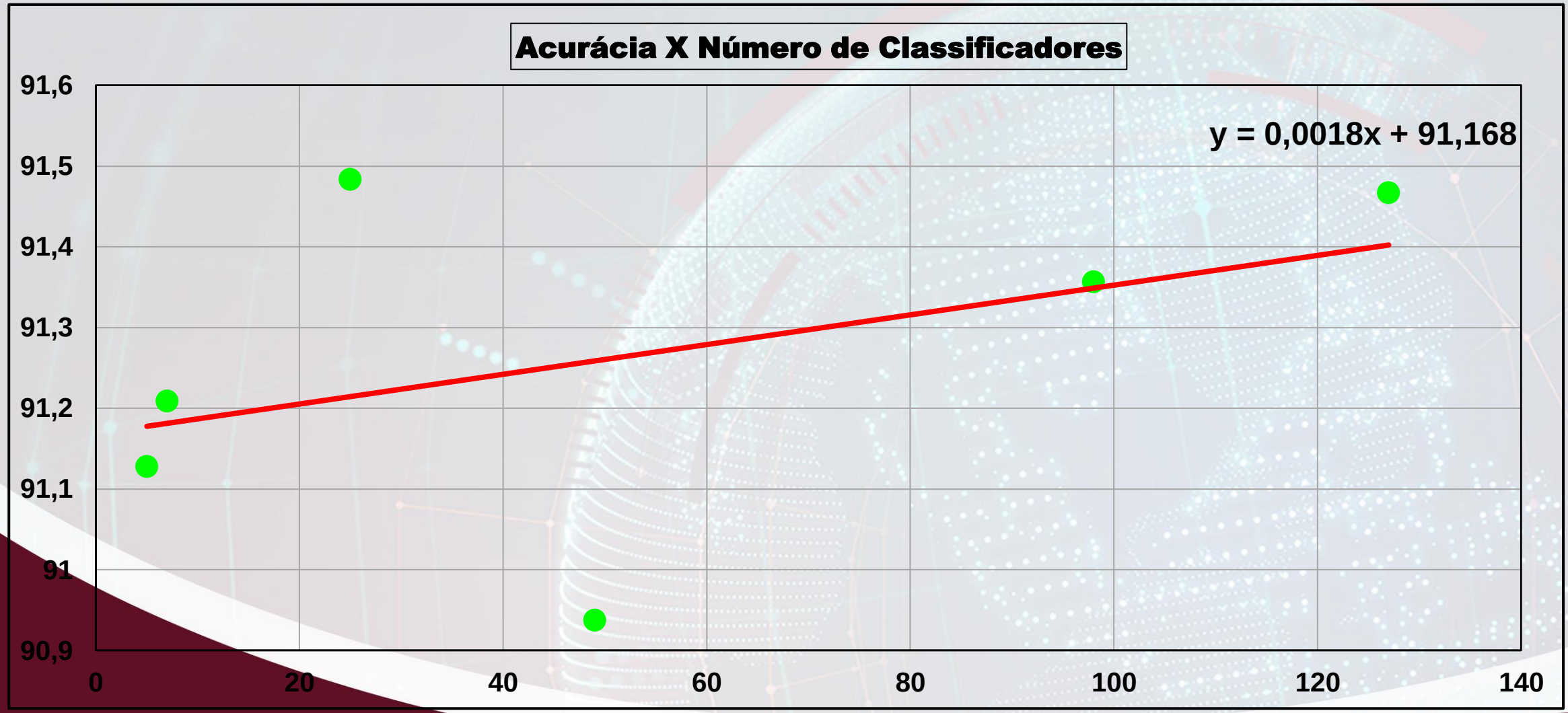
Nome	Instâncias	Ofensivo	Não Ofensivo	Língua
OffComBr2	1250	419	831	Português Brasileiro
OffComBr3	1033	201	832	Português Brasileiro
Kaggle-test	2647	693	1954	Inglês

Resultados - Determinação do número de classificadores no comitê

- **Objetivo do experimento:** Determinar o número de classificadores a utilizar nos próximos experimentos.
- **Conjunto de dados utilizado:** kaggle-test.
- **Treinamento/avaliação dos classificadores:** validação cruzada em 5 partes.

Classificadores	Acurácia Média dos Classificadores (%)	Acurácia do Comitê (%)	Tempo de Execução do Comitê (minutos)
5	90,4648	91,128	7
7	90,706	91,209	10
25	90,563	91,484	19
49	90,49	90,938	54
98	90,354	91,357	105
127	90,525	91,467	135
Média:	90,51713333	91,26383333	

Resultados - Determinação do número de classificadores no comitê



Resultados – Identificação de Linguagem Ofensiva

- **Objetivo dos experimento:** Determinar a acurácia de comitê composto por sete classificadores na tarefa de identificação de linguagem ofensiva (C-GPqSI_7).
- **Conjuntos de dados utilizados:** OffComBr2, OffComBr3 e kaggle-test.
- **Treinamento/avaliação dos classificadores:** validação cruzada em 5 partes.

Matriz de Confusão

		Resposta Esperada: Discurso de Ódio	Resposta Esperada: Não é discurso de ódio
Resposta Prevista	Discurso de Ódio	410 Verdadeiro Positivo	14 Falso Positivo
	Não é discurso de ódio	7 Falso Negativo	817 Verdadeiro Negativo

Matriz de Confusão

		Resposta Esperada: Discurso de Ódio	Resposta Esperada: Não é discurso de ódio
Resposta Prevista	Discurso de Ódio	201 Verdadeiro Positivo	14 Falso Positivo
	Não é discurso de ódio	0 Falso Negativo	818 Verdadeiro Negativo

Matriz de Confusão

		Resposta Esperada: Discurso de Ódio	Resposta Esperada: Não é discurso de ódio
Resposta Prevista	Discurso de Ódio	689 Verdadeiro Positivo	19 Falso Positivo
	Não é discurso de ódio	4 Falso Negativo	1935 Verdadeiro Negativo

Resultados – C-GPqSI$_7$ vs Hate2Vec

C-GPqSI-7 [3]

	F1	ROC-AUC	Acurácia
OffComBr2	0,98	0,97	0,98
OffComBr3	0,97	0,96	0,98
Kaggle-test	0,93	0,89	0,91

Hate2Vec [4]

	F1	ROC-AUC	Acurácia
OffComBr2	0,97	0,98	0,97
OffComBr3	0,94	0,94	0,94
Kaggle-test	0,91	0,88	0,91

Conclusões

- O C-GPqSI obtém bom resultados na tarefa de identificação de linguagem ofensiva em textos publicados em mídias sociais, superiores a abordagem recente publicada na literatura (Hate2Vec).
- A acurácia do C-GPqSI aumenta discretamente com o aumento do número de classificadores que o compõe.
- Conforme aumenta o número de classificadores, aumenta de forma linear o tempo de execução do comitê.

Referências

[1] KUMAR, R.; BHANODAI, G.; PAMULA, R.; CHENNURU, M. R. Trac-1 shared task on aggression identification: Iit (ism)@ coling'18. 2018. Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018) [...]. [S. l.: s. n.], 2018. p. 58–65.

[2] COSTA NETO, A. D. ; ADÁN COELLO, J. M. . Redes Neurais Convolucionais Aplicadas à Análise de Sentimento. In: XXII Encontro de Iniciação Científica da PUC-Campinas, 2017, Campinas. Anais do XXII Encontro de Iniciação Científica da PUC-Campinas, 2017

[3] SANCHES, C. L. S. D. T.; ADÁN COELLO, J. M. Detecção de Posicionamento em Mídias Sociais Usando Comitês de Classificadores. In: Anais do XXIV Encontro de Iniciação Científica da PUC-Campinas. 2019 set 24-26; Campinas, São Paulo.

[4] PELLE, R; ALCÂNTARA, C; MOREIRA, V. P. "A Classifier Ensemble for Offensive Text Detection", in Proceedings of the 24th Brazilian Symposium on Multimedia and the Web, 2018, p.237-243.