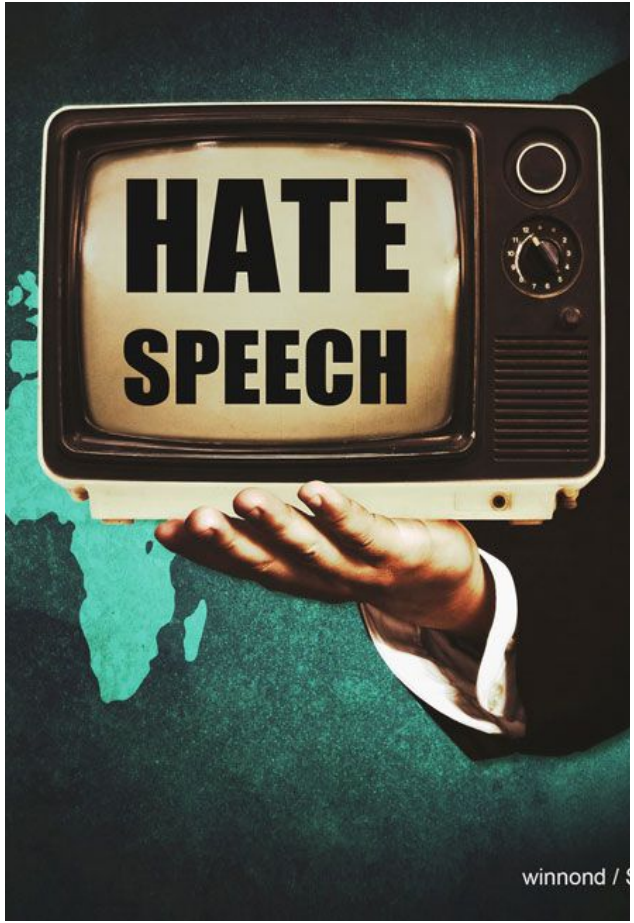
Projeto IBRA
Coleta de dados

Daniel Silva Lopes da Costa
Fábio Tamaru Nakamura
João Paulo Souza

Introdução

- **Objetivos e Tarefas**

- **Resumo Geral**
  - Descrever o método utilizado
  - Métricas utilizadas (Z2 - Score)
  - Datasets encontrados
  - Validação do método através de experimentos



Objetivos

Grupo vai ficar responsável por:

- Montar os crawlers para ingestão
- Estudar melhores formas de **anotar** os dados e **estruturar os databases.** (Eventualmente tudo será disponibilizado na internet).

Tarefas e métodos:

a) Busca de dados e Data Mining
b) Gestão de dados
c) Data Augmentation / Data Resampling / outros métodos de transformação de dados.
d) Adversarial Learning (Redes GAN): útil para gerar novos dados que desafiem bem o modelo, e também para avaliar se o modelo é bom (descobrir em que casos o modelo não é confiável).

Desenvolvemos um método de busca de dados que lembra a etapa de discriminação das redes GANs.
Validar esta ferramenta será a próxima tarefa deste grupo.

Datasets

- 14 datasets encontrados
  - 4 em português
  - 11 em inglês

- Plataformas - Inglês
  - Twitter (6)
    - 2 - Twitter id
    - 4 - Texto (1 em Json)
  - Stormfront
  - Gab (2)
  - Reddit
  - Youtube/Reddit

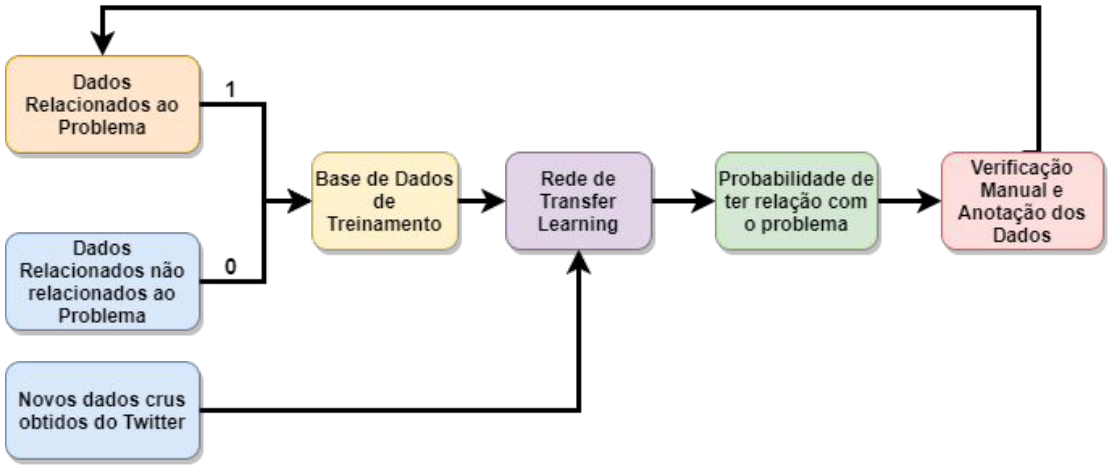
- Plataformas - Português
  - Twitter (2)
    - 1 - Twitter id
    - 1 - Texto
  - G1 - comentários
  - 55chan

Datasets

Índice	Liguagem	Plataforma	Tipo	Quantidade	Classificações	Discurso de ódio	Proporção	Link
E1	Inglês	Twitter	Texto	25296	Hate speech / offensive language / neither	4993	19,74%	https://paperswithcode.com/dataset/hate-speech-and-offensive-language
E2	Inglês	Stormfront	Texto	10945	Hate / No Hate / skip / relation	1196	10,93%	https://paperswithcode.com/dataset/hate-speech
E3	Inglês	Gab	Texto(Json)	11093	Hateful / Offensive / Normal / Undecided	5227	47,12%	https://paperswithcode.com/dataset/hatexplain
E4		Twitter		9055		708	7,82%	
E5	Inglês	Youtube/Reddit	Texto	998	Violence / Directed_vs_generalized / Gender / Race / National_origin / Disability / Religion / Sexual_orientation	433	43,39%	https://paperswithcode.com/dataset/ethos
E5'	Inglês	Twitter	Texto	5646	Abusive / offensive / hateful / disrespectful / normal / fearfull	1278	22,64%	https://paperswithcode.com/dataset/ethos
E6	Inglês	Twitter	Twitter id	28000	Variado	28000	100,00%	https://github.com/mayelsherif/hate_speech_icwsm18
E7	Inglês	Gab	Texto	11825	Binária	11169	94,45%	https://github.com/jing-qian/A-Benchmark-Dataset-for-Learning-to-Intervene-in-Online-Hate-Speech
E8		Reddit		5024		4763	94,80%	
E9	Inglês	Twitter	Twitter id	16907	Racism / Sexism / None	5348	31,63%	https://github.com/ZeerakW/hatespeech
E10	Inglês	Twitter	Texto	1797	Abusive / not abusive	368	20,48%	https://github.com/uds-lsv/lexicon-of-abusive-words
P1	Português	Twitter	Twitter id	5668	Binária	1228	21,67%	https://github.com/paulafortuna/Portuguese-Hate-Speech-Dataset
P2	Português	G1 (comentários)	Texto	1250	Binária	419	33,52%	https://github.com/rogersdepelle/OffComBR/
P3	Português	Twitter	Texto	21000	Homophobia / obscene / insult / racism / misogyny / xenophobia	9255	44,07%	https://github.com/JAugusto97/ToLD-Br

Método

- Simular o ambiente de redes sociais - usamos principalmente dados do twitter.
- Método que visa trabalhar com um número pequeno de dados rotulados **(Low Resource NLP)**



Métricas de Avaliação

O recall contabiliza o número de casos que o classificador identificou como discurso de ódio pelo total de casos de discurso de ódio. É um boa métrica pois os dados são desbalanceados tem muito caso negativo.

$$\text{Recall} = \frac{TP}{TP + FN}$$

A Precisão contabiliza fração de instâncias relevantes entre as instâncias recuperadas. Geralmente com poucos dados positivos a precisão fica alta.

$$precision = \frac{TP}{TP + FP}$$

Métricas de Avaliação

O F-beta score é uma métrica que relaciona a precisão e o recall, onde podemos conferir um peso maior ao recall.

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}.$$

Beta escolhido para os experimentos iniciais = 2

Baseline

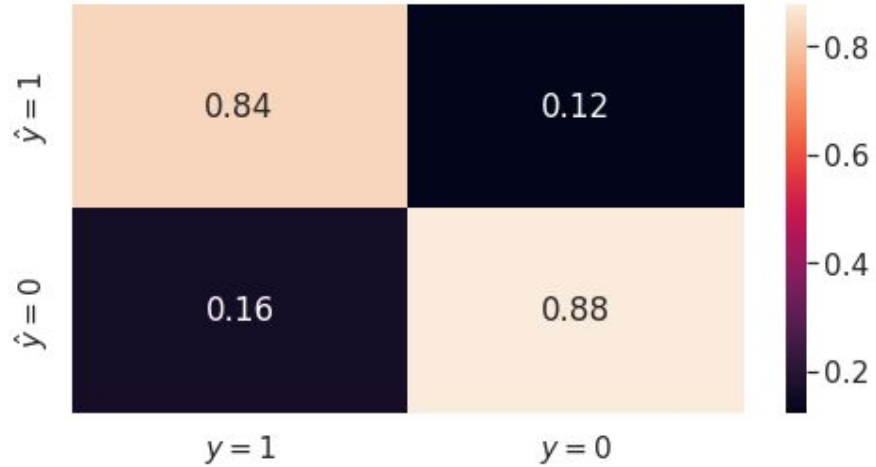
- Algoritmo simples que servirá de **Benchmark** para os modelos

- Pega as top-10 palavras mais frequentes dos dados anotados como hate speech e que não estão nos dados sem hate speech.

	Word	Frequency
0	bitch	843
1	bitches	296
2	hoes	222
3	pussy	212
4	ass	175
5	got	138
6	fuck	134
7	get	133
8	shit	121
9	nigga	113

Baseline

```
Accuracy = 18823 / 22305 (0.843892)
Recall = 15530 / 18558 (0.836836)
Precision = 15530 / 15984 (0.971597)
Fbeta Score = 0.860712
```



Treinamento:

- Usa o dataset E1:
  (**24783 linhas:** 20620 hate/offensive)
- 10% treino e 90% teste
- Os resultados foram bons, pois provavelmente o dataset foi montado pesquisando “buzz words” de ódio.

Experimento 1

- K-fold Cross-Validation no modelo
  - Chegamos em aproximações da métrica mais confiáveis
- Experimento foi feito com o E1 também
- 5 folds (80% treino e 20% teste)

	TP	TN	FP	FN	accuracy	precision	recall	Fbeta
0	699.0	25.0	127.0	120.0	0.745623	0.846247	0.853480	0.852023
1	674.0	29.0	130.0	133.0	0.727743	0.838308	0.835192	0.835813
2	672.0	29.0	151.0	120.0	0.721193	0.816525	0.848485	0.841894
3	722.0	20.0	138.0	139.0	0.728165	0.839535	0.838560	0.838755
4	696.0	31.0	159.0	143.0	0.706511	0.814035	0.829559	0.826407

	mean	standard deviation
TP	692.600000	20.537770
TN	26.800000	4.381780
FP	141.000000	13.693064
FN	131.000000	10.653638
accuracy	0.725847	0.014103
precision	0.830930	0.014629
recall	0.841055	0.009777
Fbeta	0.838979	0.009311

Experimento 1 - Próximos passos

- Usar o K-fold Cross-Validation em todos os experimentos, incluindo o Baseline
- Inverter a proporção de treino e teste **(Low Resource NLP)**

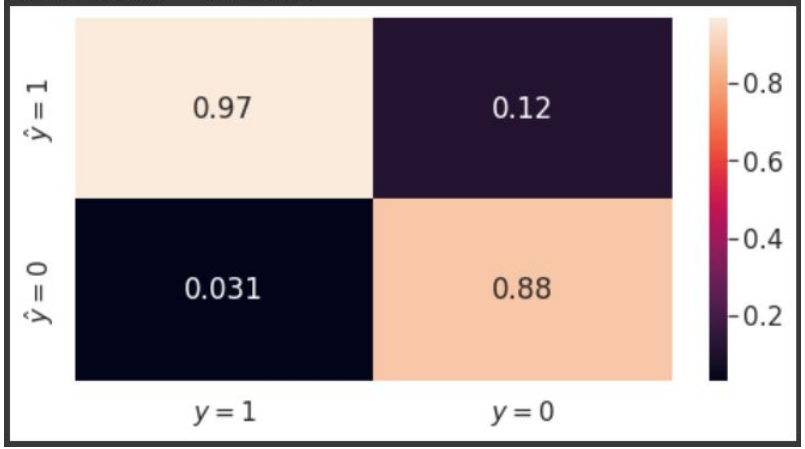
Experimento 2

- O objetivo é simular a atuação do modelo em um ambiente de redes sociais
- Uso do dataset sentiment140 com 1,600,000 de tweets com classificação por sentimento
- Etapas:
  - Treinar o modelo com uma amostra do E1
  - Pegar a outra parte do E1 e juntar com uma amostra do sentiment140
  - Testar se o modelo é capaz de identificar os tweets de discurso de ódio

Experimento 2 - Usando dados do mesmo dataset do modelo (E1)

```
TP = 10012    FP = 2688
FN =  316     TN = 19375

Accuracy = 29387 / 32391 (0.907258)
Recall = 10012 / 10328 (0.969404)
Precision = 10012 / 12700 (0.788346)
Fbeta Score = 0.926831
```

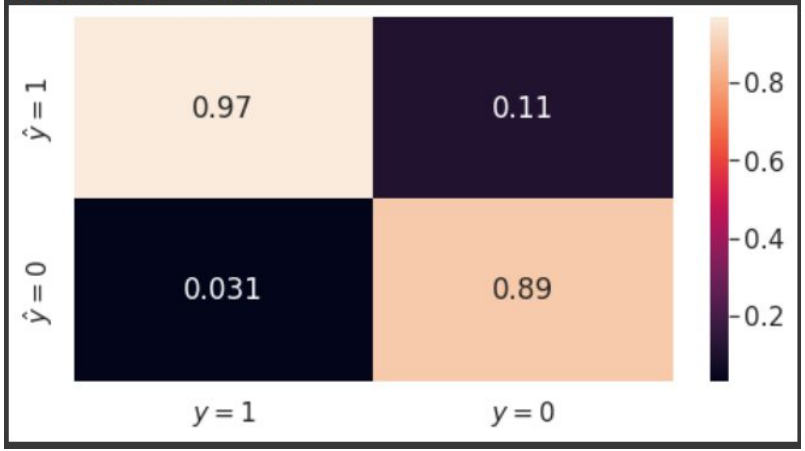


- Bons resultados no geral
- O menor indicador foi o de precisão
- Achados 10012 tweets de discurso de ódio de 10328

Experimento 2 - Usando dados do mesmo dataset do modelo (E1)

```
TP = 10012    FP = 11502
FN =  316     TN = 90561

Accuracy = 100573 / 112391 (0.894849)
Recall = 10012 / 10328 (0.969404)
Precision = 10012 / 21514 (0.465371)
Fbeta Score = 0.796804
```



- Aumentando a amostra do dataset
- Performance semelhante, porém com queda na precisão
- Achados 10012 tweets de discurso de ódio de 10328

Experimento 2 - Usando dados de outro dataset para o teste (E9)

```
TP = 1186    FP = 3544
FN = 1525    TN = 22727

Accuracy = 23913 / 28982 (0.825098)
Recall = 1186 / 2711 (0.437477)
Precision = 1186 / 4730 (0.250740)
Fbeta Score = 0.318774
```

- Recall e precisão baixos
- Alta acurácia
- Achados 1186 tweets de discurso de ódio de 2711

Experimento 2 - Próximos passos

- Treinar o modelo com um dataset diferente e observar os resultados em um terceiro dataset

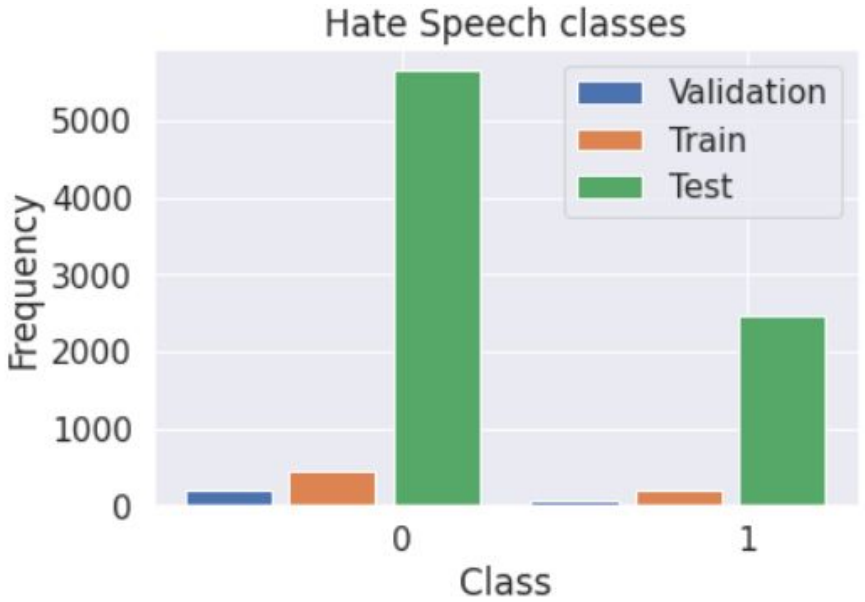
- Avaliar os dados do dataset sentiment140 e identificar se existem dados com sentimento negativo que podem ser identificados como discurso de ódio

- Treinar e testar com uma mistura de datasets

Experimento 3 - Treinar para uma subclasse

- Treinar o modelo para um subclasse específica ou retirando um subclasse
- Analisar o resultado aplicado a um teste para identificar discurso de ódio ou não.

Experimento 3 - Exemplo aplicação 1



Treinamento:

- Usa o dataset E6 - racismo(12), machismo(2711), none(6271)
- Amostra de 10% dos casos de machismos

	Treino	Test	Val
Total	628	8096	270
Sexism	196	2452	75

Experimento 3 - Exemplo aplicação 1

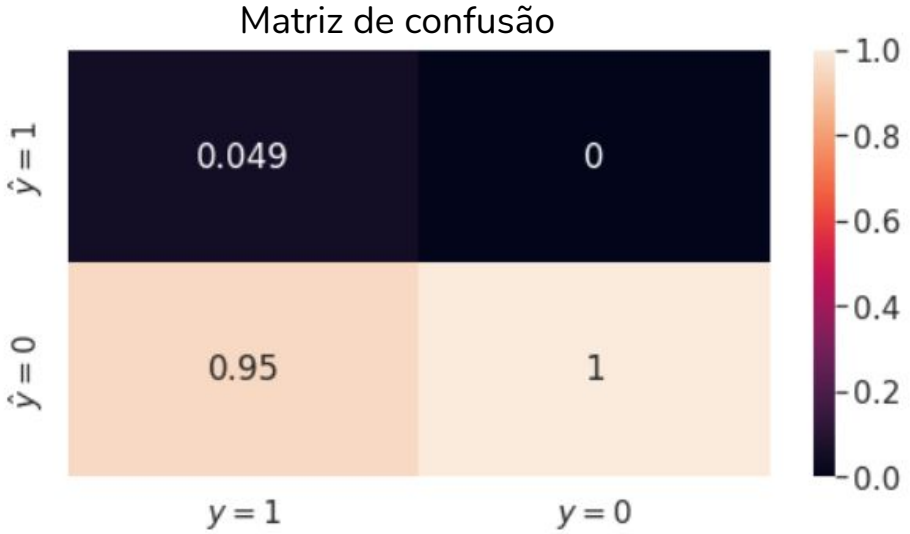
Teste 1:

Utilizando os casos de teste do dataset E6

Resultados:

```
TP =  110    FP =    3
FN = 2342    TN = 5641

Accuracy = 5751 / 8096 (0.710351)
Recall = 110 / 2452 (0.044861)
Precision = 110 / 113 (0.973451)
Fbeta Score = 0.055438
```



Experimento 3 - Exemplo aplicação 1

Teste 2:

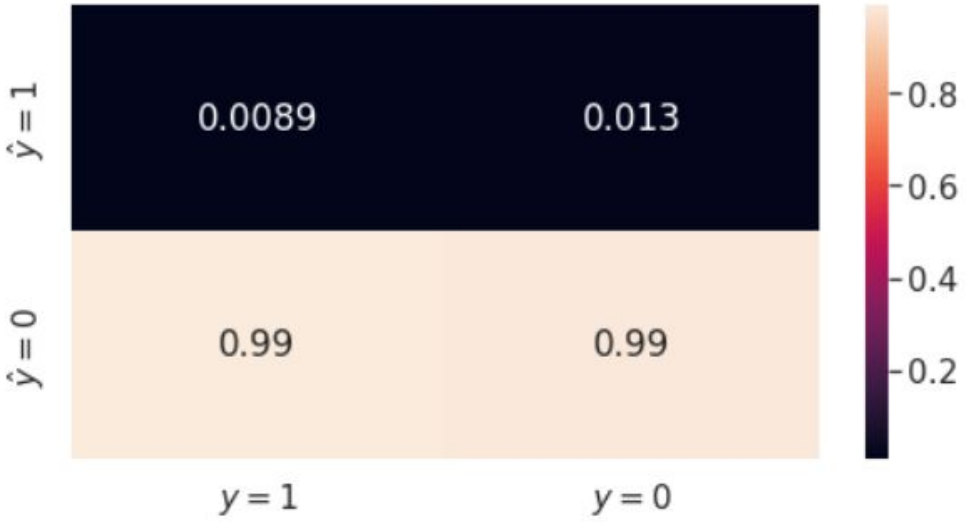
Utilizando o dataset E1

Resultados:

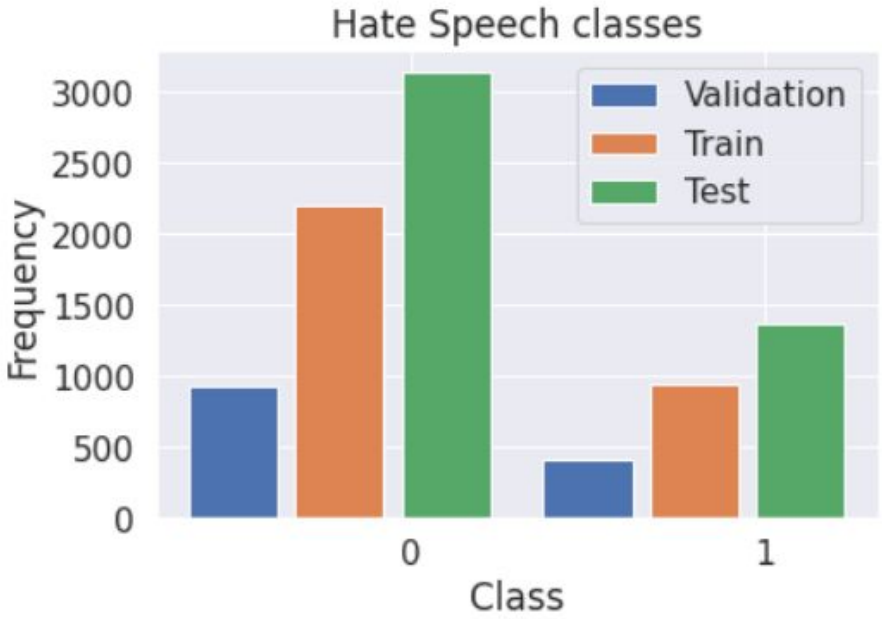
```
TP =  184    FP =   55
FN = 20436    TN = 4108

Accuracy = 4292 / 24783 (0.173183)
Recall = 184 / 20620 (0.008923)
Precision = 184 / 239 (0.769874)
Fbeta Score = 0.011122
```

Matriz de confusão



Experimento 3 - Exemplo aplicação 2



Treinamento:

- Usa o dataset E6 - racismo(12), machismo(2711), none(6271)
- Amostra de 50% dos casos de machismos

	Treino	Test	Val
Total	3143	4504	1347
Sexism	941	1368	414

Experimento 3 - Exemplo aplicação 2

Teste 1:

Utilizando os casos de teste do dataset E6

Resultados:

```
TP =  936    FP =  274
FN =  432    TN = 2862

Accuracy = 3798 / 4504 (0.843250)
Recall = 936 / 1368 (0.684211)
Precision = 936 / 1210 (0.773554)
Fbeta Score = 0.700389
```

Matriz de confusão

Experimento 3 - Exemplo aplicação 2

Teste 2:

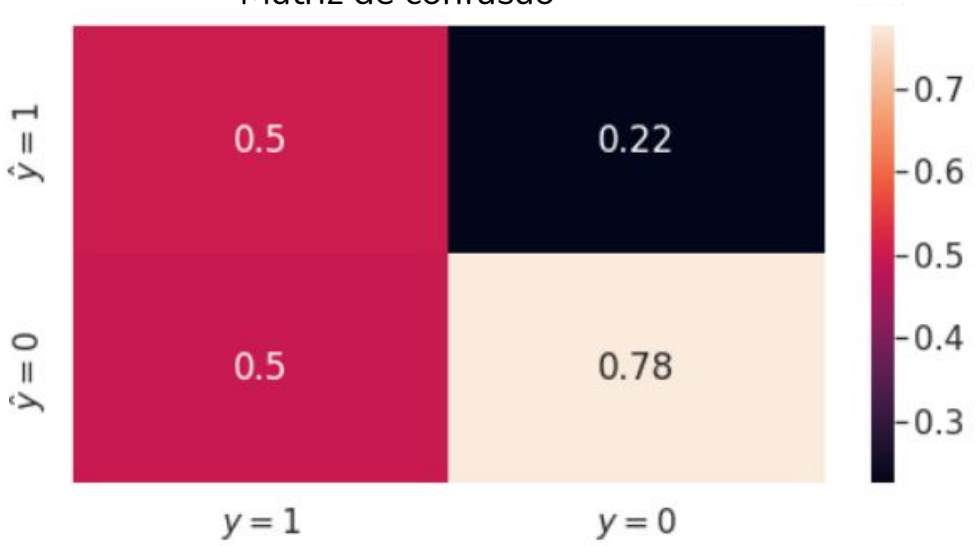
Utilizando o dataset E1

Resultados:

```
TP = 10397    FP =  935
FN = 10223    TN = 3228

Accuracy = 13625 / 24783 (0.549772)
Recall = 10397 / 20620 (0.504219)
Precision = 10397 / 11332 (0.917490)
Fbeta Score = 0.554140
```

Matriz de confusão



Experimento 3 - Próximos passos

- Aplicar o método para outros subclasses de discurso de ódio: racismo, homofobia, gordofobia.
- Usar mais datasets para treinamento.
- Fazer análises mais robustas dos resultados, treinar para diversos tamanhos de amostra e gerar mais gráficos - aprofundar análise.