Universidade Federal de Goiás
Instituto de informática
Profa Nádia Félix Felipe da Silva

Relatório: Classificação de texto - Identificação de misoginia

Aluno: Humberto Pereira Teixeira Silva e Rhuan Webster de Lourenco e Silva
Disciplina: Inteligência Computacional

Mês: Fevereiro
Ano: 2023

Universidade Federal de Goiás
Instituto de Informática
Disciplina: Inteligência Computacional

Relatório

Segundo Relatório da participação dos Alunos Humberto Pereira Teixeira Silva e Rhuan Webster de Lourenco e Silva do Curso Engenharia de Computação da Universidade Federal de Goiás, como requisito parcial para Aprovação da Disciplina Inteligência Computacional.

Alunos: Humberto Pereira Teixeira Silva e Rhuan Webster de Lourenco e Silva

Professora: Nádia Félix Felipe da Silva

Mês: Fevereiro
Ano: 2023

Conteúdo

1 Resumo

O objetivo deste relatório é mostrar e explicar todas as atividades e recursos computacionais usados para solucionar o problema que nos foi proposto na 2ª Competição da disciplina de Inteligência Computacional.

O problema em questão é a "Classificação de Texto e Identificação de misoginia". Afinal, oque é misoginia ? A palavra misoginia significa "ódio ou aversão à mulheres e meninas", sentimentos que podem se manifestar de diversas formas, incluindo violência física e psicológica, mas também a discriminação, a humilhação, a marginalização, a depreciação e a objetificação sexual da mulher. Na era digital com a aparente "liberdade"de expressão, vemos publicações em redes sociais, conteúdos misóginos que vão contra a políticas de uso da plataforma. Deste modo, será criado um classificador para identificar conteúdos misóginos a partir de textos.

Basicamente com os conjuntos de dados fornecido na competição, precisamos testar os vários algoritmos de classificações que nos foram passados na sala de aula para conseguirmos o melhor desempenho, entregando assim uma ferramenta eficiente no combate a misoginia.

2 Descrição do Conjunto de dados

O conjunto de dados é composto por textos em inglês, coletados de redes sociais e anotados manualmente, onde foram classificados como conteúdo misógino ou sem presença de misoginia. O conjunto de dados de treino estão na Figura 1.

Figura 1: Conjunto de dados de treino

	A	B	C	D	E
1	id	misogynous	text		
2	0	0	ME: WORKING REMOTELY DOING THE CHORES		
3	1	1	imgilip.com Divorce Childrens well-being Woman wh		
4	2	1	A GIRL WHO SHOWS A LOT, HAS LITTLE TO OFF		
5	3	1	Feminist: we can do everything that men do Men: m		
6	4	1	r/ConservativeMemes Posted by u/undue-influence 1		
7	5	1	This is not Feminism. iš°> >3^ This is Feminism.		
8	6	0	you have the personality of a carpeted kitchen		
9	7	1	The entire halftime show was a form of female empo		
10	8	1	hot blonde going thru a street å² No way i'n gonna al		
11	9	1	Sex Ed:"Women have one hole where the man puts		
12	10	0	IF YOU WANT A GIRLFRIEND CLOSE THE INTERI		
13	11	0	When your girlfriend is moving in with you and she a		
14	12	0	normal people: me as an intellectual:		
15	13	0	3 rules: -no wishing for death -no falling in love - no b		
16	14	1	GE MY EYES ARE UP HERE MC memecenter.con		
17	15	1	MILEKERS Hall WAWA g VICKLAR SPICKERS MT		
18	16	0	Hey, yuo want to have HETERÃ–SEX RELATION? e		
19	17	1	VIA 9GAG.COM 35.800 ONE GIRL'S CAMO IS SO		
20	18	0	Pregnant women: "Oh my god my baby is so gorgeu		
21	19	0	WHEN WOMEN ASK YOU WHY YOU ARE SINGL		
22	20	0	Make america great again Make america great Mak		
23	21	0	now everyone is cleaning as much as me huh adriar		
24	22	1	Opportunity it only passes out once		
25	23	1	MAYBE IF I RAPE HER IT'LL PUT HER IN THE MO		
26	24	0	*Bored me jerking off for the 8th time in 4 hrs* My di		

3 Descrição de atividades

Pelo conjunto de dados de treino, vemos que possui apenas um coluna, então pensamos separar uma coluna para dados de texto e outra para a classe. Nesses dados de texto criamos um dicionário com base no número de ocorrências de uma palavra no conjunto de dados e para facilitar a separação colocamos todos em minúsculo como mostrado na figura 2.

Figura 2: Criação do dicionário de palavras

```
aloneWords = textos.str.lower().str.split() #colocar todas as palavras em minúsculo
#print(aloneWords)

dicionario = set()

for lista in aloneWords:
  dicionario.update(lista) #monta um dicionario pra cada palavra do dataset

#print(dicionario)
print(len(dicionario))

wordPosition = dict(zip(dicionario, range(len(dicionario))))
#print(wordPosition)

def vetorizeCountWords(texto, wordPosition):
  vetor = [0] * len(wordPosition)
  for word in texto:
    if word in wordPosition:
      position = wordPosition[word]
      vetor[position] += 1
  return vetor

vetorText = [vetorizeCountWords(texto, wordPosition) for texto in aloneWords]
```

Após a criação do dicionário de palavras do conjunto de dados. Separamos 80% dos dados em dados treino e teste. Foi usado classificadores como Árvore de Decisão, CNN, LSTM e Classificador MLP). De todos eles, o que obteve o melhor resultado foi o classificador MLP, que na plataforma Kaggle o *score* foi de 0.74090.

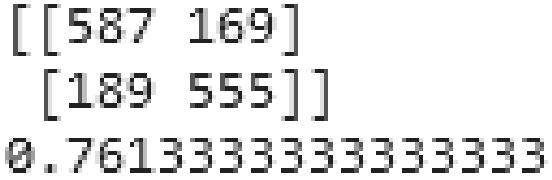
Figura 3: Classificação dos dados

```
from sklearn.neural_network import MLPClassifier

from sklearn.datasets import make_classification


X_train, X_test, y_train, y_test = train_test_split(vetorText,y, test_size=0.20, random_state=0)

#dt = DecisionTreeClassifier(max_depth=None, criterion='entropy', splitter='random', min_impurity_decrease= 0.000000001, max_
#dt = MLPClassifier(random_state=1, max_iter=300).fit(X_train, y_train)
dt = MLPClassifier(solver='sgd', alpha=1e-5, random_state=0, activation = 'logistic', learning_rate= 'adaptive')

dt.fit(X_train, y_train)

#conf = confusion_matrix(y_test, dt.predict(X_test))
accuracy = accuracy_score(y_test, dt.predict(X_test))

#clf = MLPClassifier(random_state=1, max_iter=300).fit(X_train, y_train)

#accuracy = predition.score(X_test, y_test)

#print(conf)
print(accuracy)
```

4 Análise dos Resultados

Como foi utilizados quatro classificadores, inclusive o classificador LSTM, que foi solicitado pela professora. O que obteve o melhor resultado foi o classificador MLP, que obteve o aproximado de de 76%. Analisando a matriz de confusão na figura 4 podemos ver que as ocorrências de falso negativo e falso positivo são 169 e 189 num total de 1500 dados de teste, respectivamente.

Figura 4: Resultados usando o classificador MLP



5 Trabalhos Futuros

Como essa foi segunda e última competição da disciplina irá compor a nota, não teremos um trabalho futuro a ser trabalhado.

Bibliografia

- https://scikit-learn.org/stable/ (Acessado em 19/12/2022)
- Russell, S., Norvig, P. Inteligência Artificial, Editora Campus, 2004