

Projeto final de PLN

Suicide Note Classification Using Natural Language Processing: A Content Analysis

John Pestian, Henry Nasrallah, Pawel Matykiewicz,
Aurora Bennett and Antoon Leenaars



Projeto realizado por:

Eric Shimizu

Jair Edipo

Michelle Hamada

Ricardo Gomes

Suicide

60% no mundo ↑



EUA

- segunda principal causa de morte entre pessoas de 25-34 anos
- terceira causa de morte entre as pessoas de 15-25 anos

Brasil

- 5,8 mortes por 100 mil habitantes

Hungria, Coreia

- 21 mortes a cada 100 mil habitantes

About PLN

- Principal método utilizado: Processamento da linguagem natural
 - Subárea da computação, linguística e inteligência artificial
 - Capacidade de uma máquina entender a linguagem humana natural

John Pestian

Autor principal



- Análises de características de traços e estados para identificação precoce de doenças neuro psicológicas e psiquiátricas por meio de IA

Projetos atuais:

- Fusão de pistas linguísticas, acústicas e visuais que estão sendo testadas em clínicas
- Mentoria de 5 docentes na Universidade de Cincinnati






PhD, MBA, diretor no centro de medicina computacional e professor no departamento de pediatria da UC;



Suicide Note Classification Using Natural Language Processing: a Content Analysis



- 
- 
- Objetivo do artigo
 - Base de dados
 - Tratamento das palavras
 - Métodos: KNN, SMO, Naive Bayes, Árvores de decisão.
- 
- 
- 
- 
- 
- 



Resultados do artigo escolhido (Taxa de acerto)



- Algoritmos
 - Árvores de decisão **74,4%**
 - SMO (Sequential Minimal Optimization) **70,5%**
 - Árvores de decisão de apenas um nível **66,7%**
 - Especialistas da área
 - Estagiários em psiquiatria **51,0%**
 - Profissionais da saúde **60,9%**
- 
- 
- 
- 



Base de dados

- Falta de dados
- Notícias sobre:
 - 42 suicídio
 - 16 violência
 - 10 randômico
 - 9 assédio
- Tratamento



Pré- Processamento

Tokenização Stopwords e Dicionários

- Biblioteca NLTK
- Quebra de palavras
- Remoção de pontuação
- Remoção de Stopwords
- População do dicionário global
- Criação dos dicionários de cada documento
- Big-O



Modelos de Classificação

K-Nearest Neighbors

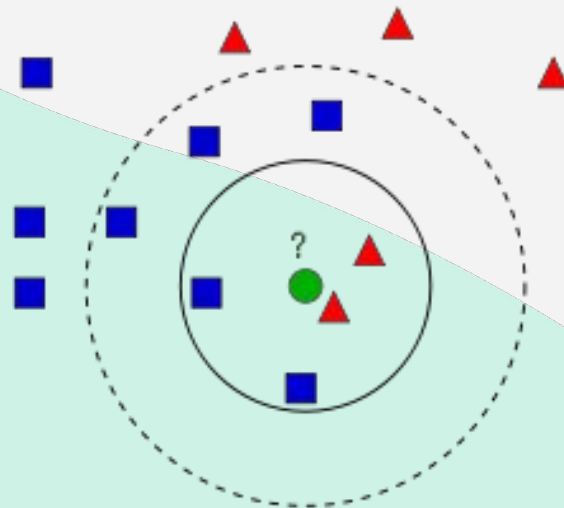
Time: $O(1)$

Space: $O(kn)$

$$\text{Euclidian} : \sqrt{\sum_{i=1}^k (x_i - y_i)^2}$$

$$\text{Manhattan} : \sum_{i=1}^k |x_i - y_i|$$

$$\text{Minkowski} : \left(\sum_{i=1}^k (|x_i - y_i|)^q \right)^{\frac{1}{q}}$$

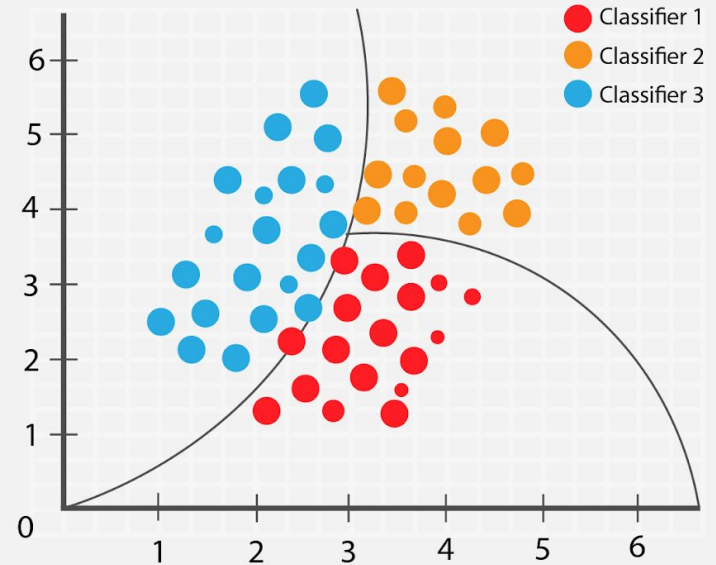


Naive Bayes

Time: $O(kn)$

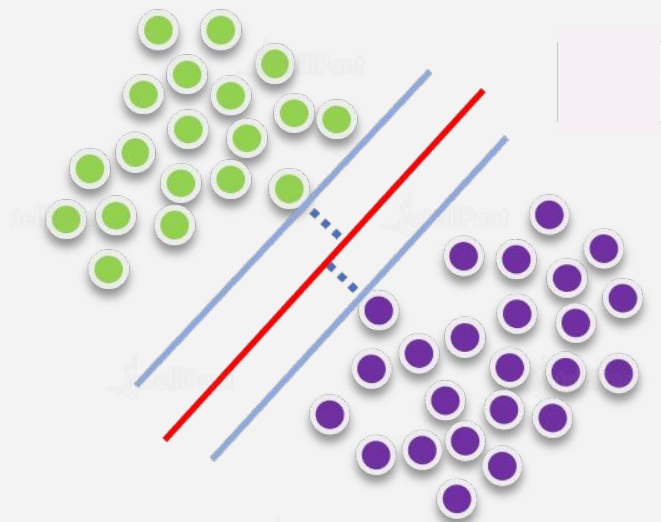
$$P(class|data) = \frac{P(data|class)P(class)}{P(data)}$$

Naive bayes classifier



Support Vector Machine

Time: $O(n^3)$



$$\text{Min } \frac{1}{2} ||w||$$

Sequential Minimal Optimization

$$L(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(x_i \cdot x_j)$$

Resultados

Acurácia: 0.5

KNN	Precision	Recall	F1-Score
Suicídio	0.67	0.29	0.40
Outros	0.44	0.80	0.57
Média	0.56	0.54	0.49

Acurácia: 0.66

NB	Precision	Recall	F1-Score
Suicídio	0.64	1.00	0.78
Outros	1.00	0.70	0.33
Média	0.82	0.60	0.56

Acurácia: 0.75

SVM	Precision	Recall	F1-Score
Suicídio	0.75	0.86	0.80
Outros	0.75	0.60	0.67
Média	0.75	0.73	0.73

Conclusão

Implementar para identificação dos assuntos relacionados ao tema



★ Análise e prevenção

Obrigado!

