Aplicações da classificação bayesiana de dados textuais à pesquisa em Psicologia Social

Hugo Cristo Sant'Anna & Maria Cristina Smith Menandro
hugo.santanna@ufes.br | cristinasmithmenandro@gmail.com

Introdução

A categorização de dados textuais na Psicologia Social é realizada manualmente ou com auxílio de programas estatísticos. Nas duas abordagens, realiza-se inferências a partir da organização e codificação dos dados quanto às frequências e associações do vocabulário, considerando que: 1) programas demandam *corpora* extensos, enquanto a análise manual pode ser feita em conjuntos reduzidos; 2) pesquisadores nem sempre compreendem as análises automáticas, ao passo que explorações manuais sucessivas dos *corpora* aumentam a confiança nas categori as propostas. A aplicação intuitiva do Teorema de Bayes neste processo pode oferecer alternativas intermediárias, explorando facilidades computacionais sem alienar os pesquisadores dos cálculos efetuados.

Objetivos

A pesquisa em andamento tem como objetivo investigar aplicações de classificadores bayesianos à categorização de dados textuais, permitindo ajustes intuitivos no processamento conforme as particularidades dos *corpora* analisados.

Método

A forma aproximada do Teorema de Bayes (Equação 1), implementada como classificador na linguagem R, foi aplicada a 61 respostas divididas em três categorias previamente definidas pelos autores. Em 100 experimentos, selecionou-se aleatoriamente 10% a 50% respostas dos *corpora* e calculou-se a probabilidade do vocabulário por categoria para formar o conjunto de treinamento. As respostas restantes foram testadas no classificador, incrementando ciclicamente a amostragem dos *corpora* em 5% até atingir precisão de categorização ≥80% em cada experimento.

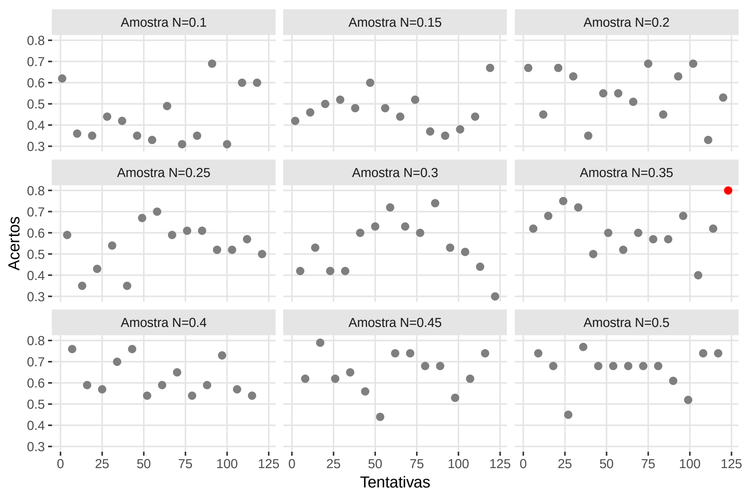

Figura 1 · Registros das tentativas e acertos do classificador com precisão ≥80% em 123 tentativas (N=61, amostra=21, teste=40)

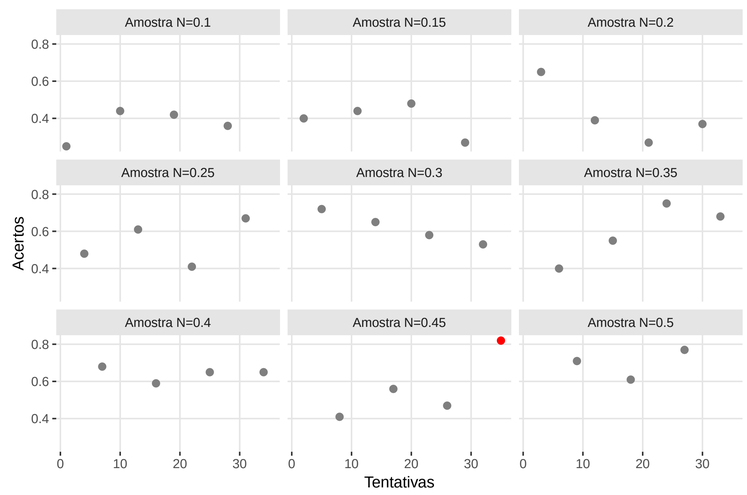

Figura 2 · Registros das tentativas e acertos do classificador com precisão ≥80% em 35 tentativas (N=61, amostra=27, teste=34)

$$pr(\text{categoria}|\text{corpus}) \approx pr(\text{categoria}) \prod pr(\text{termos}|\text{categoria})$$

Equação 1 · Forma aproximada do Teorema de Bayes

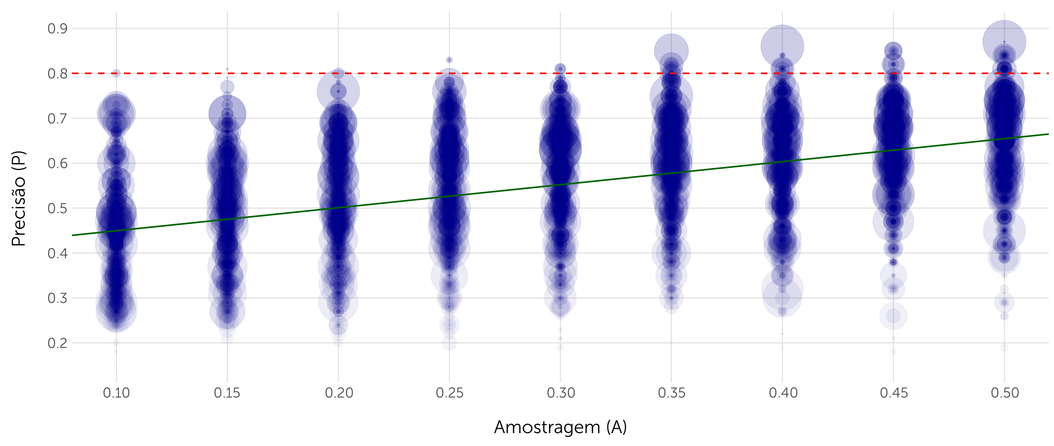

Figura 3 · Análise do desempenho do classificador durante os 100 experimentos, considerando níveis de amostragem de 10% a 50% dos *corpora* e buscando precisão ≥80% (linha tracejada). Os diâmetros indicam a quantidade de tentativas realizadas, em cada nível de amostragem, para atingir a precisão desejada. O modelo de regressão *P=0,3983+(0,5123*A)+e* está representado pela reta em verde.

Resultados esperados

O classificador atual categoriza as 61 respostas (203 *uni-grams*) com precisão ≥80% após $\bar{x}$=80 (*s*=81) tentativas, empregando conjuntos de treinamento de $\bar{x}$=26 (*s*=5) amostras. Os resultados indicam que a análise manual de 1/3 das respostas seria suficiente para que o restante dos *corpora* fosse automaticamente categorizado na precisão desejada, reduzindo consideravelmente o trabalho dos pesquisadores. Há correlação positiva moderada entre amostragem e precisão (*r*=0,50; *p*<0.01), sugerindo aplicações alternadas das abordagens manual e bayesiana no treinamento do classificador. Trabalhos subsequentes investigarão e aprimorarão o desempenho do classificador em *corpora* mais extensos e diversos.

Palavras-chave

Teorema de Bayes, Pesquisa Qualitativa, Processamento Automatizado de Dados, Análise de Conteúdo, Psicologia Social

Site do projeto

hugocristo.com.br/projetos/bayes