Classificação de Status de Saúde Mental a Partir da Análise de Comentários no Conjunto de Dados *Sentiment Analysis for Mental Health*

Caio Ribeiro de Paula
Prova Intermediária: *Natural Language Processing*
Insper
São Paulo, Brasil

I. Obtenção de Dados

Para a implementação do classificador, foi utilizado o dataset *Sentiment Analysis for Mental Health [1].*

	id	statement	status
32415	32415	and you're mine forever.	Normal
52940	52940	Panic attacks have returned - what to do? I've...	Anxiety
26466	26466	As the title says. My life is a mistake in its...	Suicidal
19551	19551	I hate warmth and going inside so its not goin...	Depression
28047	28047	I also want to set the record straight with my...	Stress

Fig. 1. Exemplificação da base de dados

II. Pipeline de Classificação

A. Pré-processamento dos Dados

O pré-processamento do dataset foi realizado para garantir a qualidade e a relevância dos dados utilizados no classificador de saúde mental. Isso incluiu a remoção de ruídos, normalização do texto para minúsculas, tokenização para dividir as declarações em palavras, e remoção de stopwords para eliminar palavras irrelevantes. Além disso, foram aplicadas técnicas de stemming e lematização para reduzir palavras a suas raízes, facilitando a identificação de padrões. As categorias de saúde mental foram convertidas para um formato numérico, permitindo que os algoritmos de aprendizado de máquina processassem os dados de forma eficaz, e estratégias para equilibrar o conjunto de dados foram implementadas, garantindo um treinamento o mais justo possível do modelo.

B. Vetorização dos Dados

A vetorização dos dados foi realizada utilizando a técnica CountVectorizer, que transforma o texto em uma matriz de contagem de termos. Essa abordagem converte cada declaração em um vetor, onde cada elemento representa a frequência de uma palavra específica no texto, permitindo assim que o classificador processe as informações textuais de maneira numérica. O CountVectorizer também facilita a eliminação de palavras irrelevantes e a consideração de apenas os termos mais significativos, o que contribui para a eficiência do modelo.

C. Aplicação do Modelo

A aplicação do modelo de classificação de saúde mental foi realizada em duas etapas principais: a classificação de tópicos e a análise de sentimentos. Inicialmente, utilizamos um classificador baseado em Multinomial Naive Bayes (MultinomialNB) para prever o estado de saúde mental das declarações. No entanto, após avaliar o desempenho do modelo, decidimos mudar para um classificador de Logistic Regression, que apresentou melhores resultados em termos de acurácia e capacidade de generalização.

A abordagem Bag-of-Words permite que o modelo identifique a ocorrência de palavras, que têm uma ligação direta com o sentimento ou o estado mental, e a frequência dessas palavras pode ser um indicativo valioso para a classificação.

III. Execução e Validação do Classificador

Usando a regressão logística pode ser útil já que esta pode produzir probabilidades de classificação que ajudam a entender melhor a incerteza associada a cada previsão, o que é particularmente relevante na análise de sentimentos, em que os detalhes no parâmetro emocional são sutis. Além disso, sua interpretação é bastante clara, permitindo que sejam identificadas quais variáveis (palavras ou termos) estão influenciando as previsões.

Dado que o dataset utilizado possui um nível de desbalanceamento, sendo que dois tipos de status relacionados à saúde mental ocupam 60% dos dados existentes, sendo ”Depression”, por exemplo, significativamente mais representada do que outras, isso leva o modelo a prever predominantemente a classe majoritária. Além disso, as características semelhantes entre algumas classes podem dificultar a distinção entre elas, enquanto a presença de ruído nos dados pode impactar negativamente as predições.

Buscando mitigar esses efeitos na análise da acurácia, esta foi obtida de forma balanceada, apresentando um valor de 70%. Fazendo-se uma análise das palavras mais relevantes no contexto do dataset, e dividindo com base em cada categoria de saúde mental, obteve-se resultados condizentes. No contexto da classe ”Bipolar”, por exemplo, a palavra ”maniac” se fez presente.

IV. Tamanho do Dataset e *Downsampling*

Testou-se proporções de 10% a 100% do total de linhas da base de dados e mediu-se a acurácia tanto no conjunto de treinamento quanto no conjunto de teste. Concluiu-se que, para os tamanhos de amostra menores, a acurácia de treinamento é bastante alta, sendo que a acurácia de teste, nesse intervalo, é relativamente baixa, indicando que o modelo pode estar se ajustando demais aos dados de treinamento, o que é um sinal de overfitting, que é quando um modelo aprende muito bem os dados de treinamento, mas falha em generalizar para novos dados.

Conforme se aumenta o tamanho da amostra, a acurácia de teste apresenta uma leve melhoria. Isso sugere que, com um maior volume de dados, o modelo se torna mais capaz de generalizar suas previsões para dados não vistos. Por outro lado, a acurácia de treinamento mostra uma leve tendência de declínio. Isso é esperado, pois à medida que mais dados são incluídos no treinamento, o modelo pode se tornar menos especializado em um conjunto de dados específico.

A combinação de um tamanho de amostra adequado e a seleção de um modelo apropriado são cruciais para obter o melhor desempenho em tarefas de classificação. Considerando a avaliação dos erros nos conjuntos de dados de treino e teste, fica claro que há espaço para aumentar a precisão do modelo ao expandir o tamanho do conjunto de dados. Essa estratégia é viável, especialmente se o aumento do conjunto de dados puder ser realizado de forma eficiente e alinhada com as necessidades do negócio e demais pontos observados ao longo deste projeto.

V. Modelos de tópicos

A fim de identificar padrões, a técnica *Latent Dirichlet Allocation* revela como determinados estados emocionais ou de saúde mental estão associados a diferentes tópicos. Isso facilita a identificação de padrões de linguagem que podem estar correlacionados com condições como depressão, ansiedade ou pensamentos suicidas.

A taxa de aceitação por tópico pode ser visualizada na imagem abaixo.

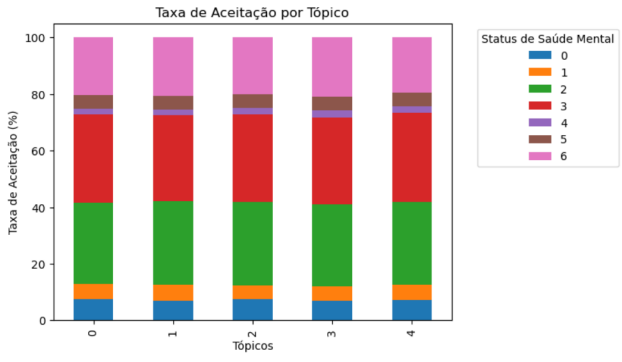

Fig. 2. Taxas de Aceitação por Tópico Encontrado

Sendo que os tópicos encontrados foram os seguintes:

- *year get go work time like would friend life depress*
- *im anxieti feel ive get day like go time start*
- *want feel like life cannot go know get fuck even*
- *feel im like dont peopl know thing think make want*
- *fuck oh http get ye world co god thank buy*

Assim, percebe-se uma disparidade em relação às porcentagens de cada tópico. Em todos os tópicos, observamos que as categorias ”Suicidal” e ”Depression” dominam, enquanto ”Normal” aparece com uma frequência menor, o que indica que os tópicos descobertos pelo LDA estão mais focados em estados emocionais graves.

VI. Conclusões

A análise de desempenho dos classificadores para cada tópico revela um padrão consistente na dificuldade em distinguir classes minoritárias, especialmente nas classes menos representadas. As acurácias gerais dos tópicos variam entre 0.68 e 0.73, indicando um desempenho razoável do modelo, mas o desbalanceamento entre as classes sugere que melhorias são necessárias na identificação de estados de saúde mental menos frequentes. A escolha de diferentes classificadores também contribuiu para a evolução do modelo, mas ainda é necessário explorar estratégias adicionais para lidar com as classes minoritárias, como técnicas de reamostragem ou ajuste de pesos. Este projeto destaca a importância de um pré-processamento robusto e de uma avaliação de desempenho equilibrada para garantir que todos os estados de saúde mental sejam representados de forma adequada nas previsões.

References

[1] S. Sarkar, ”Sentiment Analysis for Mental Health: Unlocking Mental Health Patterns through Statements,” 2024. [Online]. Available: https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health. [Accessed: Oct. 3, 2024].