TESTE 2 – RESPOSTAS

1. Como funciona o teste de hipóteses e qual é a sua finalidade na análise estatística?

O teste de hipóteses também conhecido como teste estatístico ou teste de significância consiste em uma técnica estatística para tomada de decisões sobre suposições ou afirmações feitas a respeito de uma população com base em dados amostrais.

O teste de hipóteses consiste em algumas etapas, como:

- *a formulação de hipótese, em que se tem a Hipótese Nula (H0) – supõe que não há nenhum efeito ou diferença de um grupo para outro; Hipótese Alternativa (H1 ou Ha) – suposição que compete com a Hipótese nula ou que tenta se opor a ela.*
- *escolha do nível de significância (α): indica a probabilidade de rejeitar ou não, a hipótese nula, que está associado ao p-valor (valor de probabilidade). Em outras palavras, este valor estatístico permite quantificar a evidência contra uma hipótese nula em um teste estatístico. Valores pequenos de p-valor, indica rejeitar a hipótese nula, valores grandes de p-valor, indica que não há evidências suficientes para rejeita a hipótese nula.*
- *coleta de dados e estatísticas de teste: obtém uma amostra representativa dos dados e calcula-se uma estatística de teste relevante com base nessa amostra.*
- *e tomada de decisão: que surge a partir da comparação do p-valor e o nível de significância.*

Sendo assim, o teste de hipótese permite testar suposições e hipóteses científicas com base em dados amostrais, buscando evidências estatísticas para apoiar ou refutar suas suposições; realizar inferências estatísticas; auxilia na tomada de decisão em diversas áreas e pode ser usado no controle de qualidade, avaliando se um processo ou produto atende determinadas especificações.

2. O que são redes generativas adversarias (GANs) e quais são os possíveis uso dessas redes?

As redes GANs (Redes Adversárias Generativas) são arquiteturas de redes neurais profunda. É chamada de adversária porque treina duas redes diferentes e as coloca uma contra a outra. Ela é capaz de treinar duas redes neurais para competirem entre si para gerar novos dados mais autênticos a partir de um determinado conjunto de dados de treinamento. Ou seja, uma rede gera novos dados pegando uma amostra de dados de entrada e modificando-a o máximo possível. A outra rede tenta prever se a saída de dados gerada pertence ao conjunto de dados original. Assim, a rede de previsão determina se os dados gerados são falsos ou reais.

As redes GANs podem ser usadas para gerar imagens, editar imagens para aplicações em videogames, entretenimento digital, para animação; podem gerar dados de treinamento para aplicação em outros modelos, isto é, aumentar artificialmente o conjunto de treinamento; podem ser usadas para preenchimento de dados faltantes; gerar modelos 3D a partir de fotos 2D ou imagens digitalizadas e também na geração de textos e conteúdo.

3. O que são modelos de linguagem? Qual a diferença entre LLMs e modelos de linguagem tradicionais?

Modelos de linguagem são sistemas computacionais que têm a capacidade de entender e gerar linguagem natural. Eles são projetados para lidar com tarefas relacionadas ao processamento de linguagem natural (PLN) e podem variar em complexidade, desde modelos simples de n-gramas até modelos de linguagem profundos e avançados baseados em redes neurais. A função principal de um modelo de linguagem é capturar a estrutura e a probabilidade associada às sequências de palavras em um idioma específico. Isso envolve aprender padrões gramaticais, semânticos e contextuais presentes em textos. Por outro lado, os grandes modelos de linguagem (LLM) são modelos de aprendizado profundo muito grandes que são pré-treinados em grandes quantidades de dados.

As diferenças entre os modelos de linguagem tradicionais (MLT) e as LLMs estão relacionadas:

- *ao tamanho do conjunto de dados de treinamento – as LLMs são treinadas em grandes conjuntos de dados;*
- *a sua capacidade de generalização – MLT possuem uma capacidade de generalização limitada, enquanto as LLMs possuem alta capacidade de generalização;*
- *as MLT são treinadas do zero para tarefas específicas e as LLMs, são frequentemente pré-treinadas em grandes conjuntos de dados e podem transferir conhecimento para várias tarefas;*
- *as MLT exigem pouco poder computacional para seu treinamento, enquanto as LLMs demandam por recursos computacionais significativos (como o uso de GPU);*
- *as MLT são aplicadas em algumas tarefas de PLN e as LLMs são aplicada em tarefas de geração de texto, tradução automática, resumo automático, conversação por chat, classificação de textos, geração de código, entre outros.*

4. Suponha que você tenha um conjunto de dados com três ou mais grupos para comparar e deseja determinar se há diferenças significativas entre eles. Descreva como você escolheria entre o teste ou outras técnicas estatísticas.

Antes de definir qual teste ou técnica estatística a ser utilizada é importante conhecer a definição do problema, a natureza dos dados, a distribuição dos dados, o tamanho dos grupos e os objetivos da análise. Sendo assim, poderia se optar por utilizar a análise de variância dos dados (compara as médias dos grupos), testes de comparação múltipla, PCA ou análise de cluster.

5. Qual é a importância do pré-processamento de texto em tarefas de NLP? Quais são as etapas comuns no pré-processamento de texto?

A etapa de pré-processamento de texto consiste no primeiro passo para preparar dados não estruturados para análise, tem importância crucial em NLP. Uma vez que a partir dela é realizada a limpeza e a preparação dos dados textuais antes de serem alimentados em modelos de linguagem. Geralmente, as etapas de pré-processamento envolvem: a padronização de texto (conversão de tudo para maiúsculas ou minúsculas); a remoção de caracteres especiais, pontuações,

números ou outros caracteres não relevantes; a tokenização (que quebra o texto em partes menores, os tokens); remoção de stopwords (exclusão de palavras comuns, como artigos definidos e indefinidos que não contribuem para a compreensão do contexto); lematização e stematização (que consiste na redução das palavras na sua forma base ou ao seu radical, respectivamente, ou seja, é normalização do formato); tratamento de entidades nomeadas; em tarefas de grandes volumes de dados, a redução de dimensionalidade pode ser aplicada, entre outras.

6. Descreva o processo de vetorização de texto e como modelos de linguagem como o Word2Vec ou o TF-IDF podem ser usados para representar palavras e documentos.

A vetorização de textos envolve a representação de um texto por meio de um vetor de termos. A abordagem mais comum para a vetorização de textos é atribuir a cada termo uma frequência, resultando na representação de cada documento por um vetor de termos. Cada termo no vetor possui um valor associado, indicando o grau de importância (conhecido como peso) desse termo no documento. O modelo Word2Vec (Word Embeddings) é utilizado para representar palavras em um espaço vetorial contínuo, de forma a preservar as relações semânticas entre as palavras, por outro lado, o TF-IDF (Term Frequency-Inverse Document Frequency) é utilizado para ponderar as palavras de acordo com a importância delas, as palavras que são usadas com frequência em muitos documentos terão uma ponderação mais baixa e as menos frequentes, terão uma ponderação mais alta.

7. O que é a análise de sentimento em NLP e quais são os principais métodos para realizar essa tarefa? Como você avaliaria a eficácia de um modelo de análise de sentimento?

Análise de sentimento em NLP é o processo de analisar um texto digital para determinar se o tom emocional da mensagem é positivo, negativo ou neutro. Os principais métodos para essa abordagem são:

- *abordagem com base em regras identifica, classifica e pontua palavras-chave específicas com base em léxicos predeterminados.*
- *abordagem baseada no uso de técnicas de machine learning e algoritmos de classificação de sentimentos, como redes neurais e aprendizado profundo, que envolve a criação de um modelo de análise de sentimentos e o treinamento desse modelo repetidamente com dados conhecidos para que ele possa adivinhar o sentimento em dados desconhecidos com alta precisão.*
- *abordagem híbrida, que consiste em combinar a abordagem baseada em regras e machine learning, usa recursos de ambos os métodos para otimizar a velocidade e a precisão ao derivar a intenção contextual no texto.*

A avaliação, relativa sobre a eficácia de um modelo de análise de sentimento, envolveria métricas como: 'accuracy' do modelo, complementada pelo 'recall', 'precision' e 'F1-score', pois esses itens mostram a proporção de predições corretas em relação ao total de predições e permite também analisar se o modelo tem suas classes balanceadas; análise exploratória sobre os resultados apresentados, a fim de

definir algum insight sobre o comportamento do modelo e a técnica de validação cruzada.

8. Qual é a diferença entre a classificação de texto e o agrupamento (clustering) de texto em NLP? Em que situações cada um é mais apropriado?

A classificação de texto tem como objetivo atribuir uma ou mais categorias predefinidas (rótulos) a um texto com base no conteúdo do texto. Sendo um problema de aprendizado supervisionado, isto é, o modelo de aprendizado de máquina é treinado em um conjunto de dados de documentos rotulados. O conjunto de dados de treinamento deve incluir documentos de cada classe que o modelo deverá ser capaz de identificar. Enquanto, o agrupamento de texto (clustering) tem como objetivo agrupar textos semelhantes com base em características ou padrões semelhantes, mas sem rótulos predefinidos. Consistindo, em um problema de aprendizado não supervisionado, o que significa que o modelo de aprendizado de máquina não é treinado em um conjunto de dados de dados de documentos rotulados. O modelo de aprendizado de máquina deve aprender a identificar similaridades entre os documentos por si só.

A classificação de texto, em geral, é aplicada em filtro de spam (identificação de e-mails indesejados), reconhecimento de tópicos (classificação de notícias com base em tópicos específicos) e análise de sentimento (identificação do sentimento positivo, negativo ou neutro de um texto. O agrupamento de texto é aplicado em problemas de segmentação de clientes (agrupamento de clientes com base em seus interesses ou comportamentos), filtragem de conteúdo (agrupamento de conteúdo semelhante para facilitar a navegação) e pesquisa de texto (agrupamento de documentos de texto relacionados para facilitar a recuperação de informações).

9. Explique o conceito de reconhecimento de entidades nomeadas (NER) em NLP e suas aplicações práticas.

O reconhecimento de entidades nomeadas (REN) é uma técnica de processamento de linguagem natural (NLP) que identifica e classifica entidades nomeadas em um texto. Uma entidade nomeada é uma expressão que pode ser interpretada como um referente específico, como nomes de pessoas, organizações, locais, datas, valores monetários, porcentagens, um evento, entre outros. Sendo útil para extração de informações, filtro de spam, pesquisa de texto, chatbots e outros.

10. Como você lidaria com problemas de desequilíbrio de classe em tarefas de classificação de texto em NLP? Quais estratégias seriam eficazes?

O desequilíbrio de classes consiste no fato em que uma classe é significativamente mais prevalente do que as outras. Há inúmeras técnicas para resolver tal problema, como: reamostragem de dados – que consiste em alterar o tamanho do conjunto de dados (sobreamostragem ou subamostragem); peso de classes (Essa estratégia envolve atribuir pesos diferentes às amostras de cada classe durante o treinamento do modelo); uso de ajuste de limiar de decisão (para equilibrar o 'precision' e o 'recall'); entre outros. É importante ressaltar que a adoção da estratégia dependerá do domínio específico do problema e do conjunto de dados de treinamento.