

QUESTÃO 1:

QUANTO AOS FUNDAMENTOS, OS MODELOS BASEADOS EM CONTAGEM SÃO OS MAIS SIMPLES, COM UMA MATEMÁTICA SIMPLES. BASEIAM-SE ^{EM} VÍCIO UM CORPUS (CONJUNTO DE TEXTOS), UM VOCABULÁRIO (CONJUNTO DE PALAVRAS OU TOKENS), APÓS UM PRE-PROCESSAMENTO DAS PALAVRAS DO CORPUS NOS TOKENS, EM UMA CONTAGEM DO NÚMERO DE VEZES CADA TOKEN DO VOCABULÁRIO APARECE NO CORPUS. PARA ISSO, CADA ITEM DO VOCABULÁRIO FUNCIONA COMO UM "ONE-HOT ENCODING". ISSO GERA MATRIZES ~~ESPARSAS~~ ACTAMENTE ESPARSAS.

JÁ OS MODELOS PROBABILÍSTICOS FUNCIONAM, APÓS O PÓS-PROCESSAMENTO, ATRIBUINDO UMA PROBABILIDADE A CADA TOKEN DE SER O PRÓXIMO APÓS UMA SEQUÊNCIA DE TOKENS DE INPUT. ISSO PODE SER GENERALIZADO PARA UMA SEQUÊNCIA DE TOKENS DE SAÍDA (OUTPUT), DADA UMA SEQUÊNCIA DE TOKENS DE ENTRADA (INPUT).

JÁ OS MODELOS BASEADOS EM EMBEDDINGS,
PASSAM PELA CONSTRUÇÃO DE UM ESPAÇO LATENTE
QUE CODIFICA EM UMA DIMENSÃO REDUZIDA,
EM RELAÇÃO À CARDINALIDADE DO CONJUNTO
DE TOKENS ASSOCIADO AO VOCABULÁRIO,
A REPRESENTAÇÃO VETORIAL DE CADA TOKEN
OU UMA "FEATURE" (CARACTERÍSTICA) QUE
COVARIAS
SE PODE CONSTRUIR A PARTIR DOS TOKENS.
CADA TOKEN DE INPUT ^{OU FEATURE BASEADA NELE} É CONVERTIDO NUM
VETOR DO EMBEDDING VIA UM "ENCODER"
(CODIFICADOR) E O "DECODER" ASSOCIADO
(DECODIFICADOR)
FAZ O PAPEL INVERSO, LEVANDO A REPRESENTAÇÃO
VETORIAL DO EMBEDDING NOVAMENTE
TE PARA ~~ESPAÇO DE~~ UM ELEMENTO DO
CONJUNTO DE TOKENS. MODELOS BASEADOS
EM EMBEDDINGS PODEM POSSUIR MÚLTIPLAS
CAMADAS DE CODIFICAÇÃO E DE DECODIFICAÇÃO.

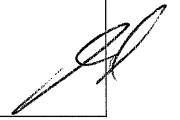
SOBRE AS VANTAGENS DE CADA ABORDAGEM;

- MODELOS DE CONTAGEM SÃO SIMPLES MATEMATICAMENTE, VÃO SER DETERMINÍSTICOS DADOS OS MESMOS CORPUS E VOCABUCÁRIO E FIXADOS SEUS PARÂMETROS. ISSO PERMITE EXPLICABILIDADE DO MODELO. JÁ OS MODELOS PROBABILÍSTICOS ~~SO~~ POSSUEM UMA MATEMÁTICA ASSOCIADA MAIS COMPLICADA, MUITAS VEZES REQUERENDO USO DE APROXIMAÇÕES DE DISTRIBUIÇÕES PROBABILÍSTICAS, MAS QUE SÃO DE BAIXO CUSTO COMPUTACIONAL PARA O HARDWARE MODERNO. SÃO PORTANTO ESTREMAMENTE RÁPIDOS. FINALMENTE, OS MODELOS BASEADOS EM EMBEDDINGS PODEM SER

TREINADOS EM CORPUS GIGANTESCOS E SEUS (DE TÓPICOS DIVERSOS)

MODELOS JÁ TREINADOS PODEM SER POSTERIORMENTE ESPECIALIZADOS, ATRAVÉS DE "FINE-TUNING", PARA UM TÓPICO ESPECÍFICO.

ISSO PERMITE QUE O CUSTO COMPUTACIONAL DO TREINAMENTO SOBRE O DATASET DE MÚLTIPLOS TÓPICOS SEJA MITIGADO, E SOB BAIXO CUSTO COMPUTACIONAL, REFINADO PARA OS TEXTOS DE INTERESSE. É POSSÍVEL TAMBÉM UTILIZAR DIRETAMENTE O MODELO TREINADO SOBRE MÚLTIPLOS TÓPICOS DIRETAMENTE PARA PREDIÇÃO EM UM TÓPICO ESPECÍFICO, APOESAR QUE FREQUENTEMENTE TAL ABORDAGEM PRODUZ RESULTADOS DE QUALIDADE INFERIOR ÀQUELA JÁ REFERIDA VIA "FINE-TUNING". OUTRO BENEFÍCIO É QUE JÁ EXISTEM MÚLTIPLOS MODELOS GRATUITOS DISPONÍVEIS, EVITANDO ASSIM O CUSTO DE TREINAMENTO. DO PONTO DE VISTA DE QUALIDADE DAS PREDIÇÕES, MODELOS DE EMBEDDINGS COSTUMAM PRODUIR MELHORES RESULTADOS DO QUE OS DITOS PROBABILÍSTICOS, JÁ QUE PODEM RECORRER A TÉCNICAS COMO TRANSFORMERS, ~~PARA~~ VIA MECANISMO DE "SELF-ATTENTION", PARA APRENDER CORRELAÇÕES ENTRE ELEMENTOS DA SEQUÊNCIA DE TOKENS COM MAIOR "RANGE" QUE OS ~~PROBABILÍSTICOS~~ PROBABILÍSTICOS. TRANSFORMERS TAMBÉM PERMITEM O PROCESSAMENTO DE TOKENS EM PARALELO.

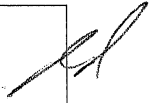


QUANTO ÀS DESVANTAGENS OU LIMITAÇÕES:

PELAS MATRIZES ESPARSAS ASSOCIADAS AOS MODELOS DE CONTAGEM, SEU CUSTO COMPUTACIONAL FICA INVIÁVEL PARA UM VOCABULÁRIO GRANDE DE TOKENS. ISSO REDUZ A CAPACIDADE DE REFINAMENTO DOS MODELOS EM DETALHES DOS TÓPICOS ABORDADOS, OU APENAS PODER SER EFETIVAMENTE UTILIZADOS EM TÓPICOS ESPECIALIZADOS. JÁ OS MODELOS PROBABILÍSTICOS SOFREM EM SEREM NÃO-DETERMINÍSTICOS (INFLUENCIANDO PORTANTO SUA INTERPRETABILIDADE E AS CONSEQUÊNCIAS DAI AINDAS, COMO QUESTÕES DE ÉTICA E GOVERNANÇA, ENTRE OUTRAS) E ACUMULARÃO ERRO QUANDO UTILIZADOS PARA PREDIÇÃO NUMA SEQUÊNCIA DE TOKENS DESAÍDA, ISTO É, UM TOKEN PREDITO FUNCIONARÁ COMO INPUT PARA PREDIÇÃO DE UM FUTURO TOKEN, E OS ERROS ASSOCIADOS SE ACUMULAM.

QUANTO AOS MODELOS DE EMBEDDING, ESSES POSSUEM ALTÍSSIMO CUSTO COMPUTACIONAL PARA TREINAMENTO, ^{FREQUENTEMENTE} REQUERENDO TÉCNICAS DE PROCESSAMENTO DISTRIBUÍDO EM GPUS, E PORTANTO ALTA DEMANDA ENERGÉTICA, PARA QUE PRODUZAM AS JÁ DISCUTIDAS ~~REDES~~ VANTAGENS.

DO PONTO DE VISTA DE APLICAÇÕES, OS MODELOS DE CONTAGEM VÃO SER EFETIVOS PARA VOCABULÁRIOS PEQUENOS, FREQUENTEMENTE UTILIZADOS EM TAREFAS DE CLASSIFICAÇÃO. OS MODELOS PROBABILÍSTICOS JÁ PODEM SER UTILIZADOS EM TAREFAS DE TRADUÇÃO DE IDIOMAS, POR EXEMPLO. FINALMENTE, OS MODELOS BASEADOS EM EMBEDDING, PODEM SER UTILIZADOS EM QUALQUER PROBLEMA DE SEQUÊNCIA DE TOKENS. DO PONTO DE VISTA DE MODELAGEM, TEXTOS, SÉRIES TEMPORAIS, ÁUDIOS DE FALA, ~~...~~ ETC, PODEM SER CONVERTIDOS EM TOKENS E, ATRAVÉS DE ENCODERS, POSSUIR SUA REPRESENTAÇÃO NUM ESPAÇO LATENTE. AS APLICAÇÕES ~~...~~ PODEM UTILIZAR O PAR ENCODER-DECODER, COMO PARA TRADUÇÃO DE IDIOMAS, APENAS O ENCODER- PARA TAREFAS DE CLASSIFICAÇÃO, E APENAS O DECODER



PARA USOS COMO CHAT MODELS, GERADORES
DE TEXTO. DE FATO, GERADORES DE TOKENS,
QUE PODEM SER ESSES OS JÁ DITOS TEXTOS,
MAS TAMBÉM VALORES DE SÉRIES TEMPORAIS,
OU UMA CODIFICAÇÃO DE FONEMAS NO CASO
DE ÁUDIO. AS ÁREAS DE APLICAÇÃO VÃO ENTÃO
DESDE TEXTOS, USOS MÉDICOS, EM FINANÇAS,
ETC.

QUESTÃO 3:

AS ETAPAS DE PRÉ-PROCESSAMENTO PERMITEM A DESCOBERTA DO CONHECIMENTO VIA AVALIAÇÃO DE PADRÕES DA SEGUINTE MANEIRA:

- A REMOÇÃO DE OUTLIERS OU SEU TRATAMENTO ELIMINA EXEMPLOS QUE FOGEM À DISTRIBUIÇÃO DO RESTO DOS DADOS;
- O POSSÍVEL COMPLETAMENTO DE VARIÁVEIS FALTANTES EM EXEMPLOS DO CONJUNTO DE DADOS PERMITE AJUDAR O MODELO NA SUA CONSTRUÇÃO. TAL PODE SER FEITO VIA MÉDIA DA DISTRIBUIÇÃO DOS OUTROS EXEMPLOS OU MESMO CONSTRUINDO UMA NOVA VARIÁVEL QUE IDENTIFICA A AUSÊNCIA DE VALOR NO ATRIBUTO ORIGINAL, DENTRE OUTRAS MANEIRAS
- O MAPEAMENTO DOS VALORES NUM INTERVALO $[0, 1]$, $[-1, 1]$, OU MESMO VIA ESCALAMENTO PELA DISTRIBUIÇÃO NORMAL (QUANDO SE IDENTIFIQUE O SEU USO A PARTIR DA DISTRIBUIÇÃO DOS DADOS DE EXEMPLO) PERMITE A ELIMINAÇÃO

DA DIFERENÇA DE ESCALA ENTRE OS ATRIBUTOS, EVITANDO QUE O MODELO ATRIBUA PESOS INCONSISTENTES A UM DETERMINADO ATRIBUTO. ISSO, POSTERIORMENTE, EVITA A FALSAS IDENTIFICAÇÕES DA CONTRIBUIÇÃO DE CADA ATRIBUTO PARA O RESULTADO DO MODELO TREINADO E POSSÍVEIS GENERALIZAÇÕES QUE SE FAZEM COM ELE.

JÁ NA SELEÇÃO DE ATRIBUTOS, PERMITE-SE QUE:

- ELIMINEMOS ATRIBUTOS ALTAMENTE CORRELACIONADOS (CASO CONTRÁRIO, O MODELO TEM DIFICULDADE NA ATRIBUIÇÃO DE PESOS). APESAR DE ESSENCIAL PARA O TREINAMENTO DOS MODELOS, A SELEÇÃO DO ATRIBUTO QUE SERÁ UTILIZADO NO TREINAMENTO DO MODELO, ELIMINA AQUELES ALTAMENTE CORRELACIONADOS, TRADUZINDO-SE EM POSSÍVEIS PROBLEMAS DE INTERPRETABILIDADE;

[Handwritten signature]

- TÉCNICAS DE CONSTRUÇÃO DE MODELOS COM ADIÇÃO DE UM ATRIBUTO POR VEZ PODEM AJUDAR NO DESEMPENHO PREDITIVO E ESTABILIDADE DOS RESULTADOS.
- O USO DE MÚLTIPAS ~~TECNICAS~~ TÉCNICAS, COMO REDES NEURAIS, RANDOM FOREST, E MESMO MODELOS LINEARES PODEM CONTRIBUIR PARA AUXILIAR NA SELEÇÃO DE ATRIBUTOS COMUNS AOS MODELOS, ANTES DO TREINAMENTO COM UMA DETERMINADA TÉCNICA SOBRE O CONJUNTO DE ATRIBUTOS SELECIONADOS, e então POSTERIOR ANÁLISE DA INFLUÊNCIA DE CADA ATRIBUTO PARA O MODELO FINAL.
- O FATO DOS MODELOS SEREM CAPAZES DE IDENTIFICAR RELAÇÃO ENTRE OS ATRIBUTOS DEVE AINDA SER INTERPRETADO COM CAUTELA JÁ QUE CORRELAÇÕES ESTATÍSTICAS NÃO IMPLICAM CAUSALIDADE. E QUE PADRÕES ESPURIOS PODEM APARECER. ~~...~~



- TAIS TÉCNICAS AJUDAM A MITIGAR OS PROBLEMAS DE GENERALIZAÇÃO E IDENTIFICAR POSSÍVEIS VIÉSSES DE SELEÇÃO EM ETAPAS ANTERIORES.

QUESTÃO 2 :

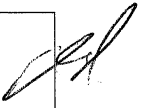
- PARA A MITIGAÇÃO DE VIESES SERIA PRECISO COMPREENDER QUE ATRIBUTOS (FEATURES) OU COMBINAÇÃO DE ATRIBUTOS LEVAM A TAIS VIESES. SE A ELIMINAÇÃO DESTES FOR OSUFICIENTE PARA ELIMINAR OS VIESES, MANTENDO A CAPACIDA DE PREDITIVA DO MODELO, TEMOS UMA SOLUÇÃO. DO CONTRÁRIO, PODE-SE CONSTRUIR NOVOS ATRIBUTOS, DE FORMA A AINDA ELIMINAR OS VIESES E QUE O SISTEMA AINDA SEJA COMERCIALMENTE VIÁVEL.

- PARA GARANTIR A REPRODUTIBILIDADE É PRECISO VERIFICAR SE: EXISTEM CARACTERÍSTICAS NÃO-DETERMINÍSTICAS NO MODELO, SE MESMO

EM ~~OS~~ MODELOS PROBABILÍSTICOS SE UTILIZA A POSSIBILIDADE DE FIXAÇÃO DE SEMENTE ALEATÓRIA DO GERADOR DE NÚMEROS ALEATÓRIOS, SE A GERAÇÃO DE NÚMEROS ALEATÓRIOS COM SEMENTE

FIXADA DE FATO GERA A MESMA SEQUÊNCIA EM HARDWARES DIFERENTES (ISSO PODE VARIAR CONFORME FABRICANTE DE PROCESSADOR), SE A DISTRIBUIÇÃO ESTATÍSTICA DOS ATRIBUTOS É A MESMA PARA DIFERENTES AMBIENTES - E SE NÃO O FOR, QUAL A RAZÃO E COMO CONSERTAR (SE POSSÍVEL) TAL QUESTÃO, SE DE FATO OS MESMOS ATRIBUTOS ESTÃO SENDO UTILIZADOS NOS DIFERENTES AMBIENTES;

- OUTRA POSSIBILIDADE, CASO AS PROPOSTAS ACIMA NÃO SEJAM SUFICIENTES, É AVALIAR A CONSTRUÇÃO DE DIFERENTES MODELOS PARA CADA AMBIENTE, O QUE SÓ SERIA VIÁVEL SE O NÚMERO DE AMBIENTES FOR PEQUENO (O QUE NÃO É EXPLICITADO NO CENÁRIO APRESENTADO);



- A PASSAGEM À EXPLICABILIDADE DO MODELO SÓ PODERIA SER FEITA CASOS OS PONTOS ANTERIORES, DE REPRODUTIBILIDADE E IDENTIFICAÇÃO DE FATORES FOSSE CONCLUÍDO. SEM REPRODUTIBILIDADE, É IMPOSSÍVEL EXPLICAR A RELAÇÃO DA SAÍDA DO MODELO COM OS VALORES DE ENTRADA. E SEM A IDENTIFICAÇÃO DE FEATURAS/ATRIBUTOS, NÃO PODEMOS SABER SEUS PESOS E PORTANTO SUAS IMPORTÂNCIAS RELATIVAS NO RESULTADO FINAL.

- MODELOS LINEARES OU MODELOS BASEADOS EM ÁRVORES DE DECISÃO PERMITEM UMA RÁPIDA IDENTIFICAÇÃO DA IMPORTÂNCIA DE FATORES. CASO O MODELO SEJA ALTAMENTE NÃO-LINEAR É DE DIFÍCIL IDENTIFICAÇÃO DE FATORES A-PRIORI, COMO NO CASO DE GRANDES MODELOS DE REDES NEURAIS

PROFUNDAS, PODE-SE RECORRER A TÉCNICAS MODERNAS DE IDENTIFICAÇÃO DE ATRIBUTOS A POSTERIORI.

- PARA CUMPRIR COM AS DETERMINAÇÕES DE ~~AMBAS~~ AMBAS LEIS DE PROTEÇÃO DE DADOS, OS CIENTISTAS DE DADOS PRECISAM GARANTIR ANONIMIZAÇÃO DE DADOS DOS USUÁRIOS E A INVESTIGAÇÃO SOBRE SE DADOS COMO IDADE E/OU SEXO PRODUZEM VIÉSSES ^{IL} ENTRE OUTROS

QUE DEVEM SER ELIMINADOS. GESTORES MUITAS VEZES DEVEM SER RESPONSÁVEIS POR ISOLAR OS AMBIENTES EM QUE HÁ O PRE-PROCESSAMENTO E A ANONIMIZAÇÃO DOS DADOS DAQUELE EM QUE OUTROS CIENTISTAS DE DADOS REALIZAM A SELEÇÃO DE FATORES/ATRIBUTOS/FATURAS E O TREINAMENTO DE MODELOS;

- O RISCO DA UTILIZAÇÃO DAS PROPOSTAS APRESENTADAS É QUE INVIABILIZEM A CONSTRUÇÃO DE BONS MODELOS A SEREM UTILIZADOS DE FORMA COMERCIAL;
- O BENEFÍCIO VÊ-SE PARA OS UTILIZADORES, QUE VÊM MITIGADOS, E SE POSSÍVEL ELIMINADOS, FATORES QUE PRIVILEGIAM DETERMINADOS SETORES/GRUPOS/CLASSES SOCIAIS, ~~PODE~~ O BENEFÍCIO PARA A EMPRESA TAMBÉM É NÃO SE VER ENCONTRAR NUM STATUS LEGAL QUE POSSA CRIAR CONFLITOS JURÍDICOS FUTUROS E PORTANTO FLUXOS DE CAIXA INESPERADOS, QUE PODEM AFETAR A SUA SOBREVIVÊNCIA.
- A PROPOSTA AQUI APRESENTADA MOSTRA COMO ELIMINAR AS QUESTÕES RELEVANTES A AMBAS AS LEIS. TAMBÉM PERMITEM IDENTIFICAR OS FATORES QUE CRIARAM OS PROBLEMAS APRESENTADOS NO CENÁRIO, JUNTAMENTE COM A SOLUÇÃO TÉCNICA PELA CRIAÇÃO DE VÁRIOS MODELOS PARA CADA AMBIENTE OU PELA IDENTIFICAÇÃO DA CAUSA DA FALTA

DE REPRODUTIBILIDADE, AINDA PERMI-
TINDO O USO DE UM SÓ MODELO COM
OS ATRIBUTOS CORRETOS DE FORMA A
ELIMINAR OS VIÉSSES.