Universidade Federal do Rio Grande do Sul

Instituto de Informática

Programa de Pós-Graduação em Computação

CMP263 - Aprendizagem de Máquina

Professora: Mariana Recamonde Mendoza

Alunos:

Augusto Giani - 00264890
Fernando Barreto Farias - 00099761

Título: Pré-processamento de dados

1. Análise da ideia inicial

Inicialmente, se colocou a hipótese de usar um dataset proprietário, com dados relativos a denúncias recebidas pelo Ministério Público do Trabalho através do seu sítio web, disponível a pessoas que sintam estar em alguma situação que fira seus direitos trabalhistas e que atinja uma coletividade de trabalhadores (não somente seus direitos individuais). A escolha foi feita em função de um dos integrantes ser servidor do órgão (Fernando Farias), e este, além de ter o acesso aos dados, possuir experiência no problema em questão: dado o texto de uma denúncia, classificá-la, de forma automática, na coordenadoria correspondente (cada coordenadoria é especializada em um assunto). Dessa classificação depende o direcionamento a um grupo de procuradores, que atua em uma determinada coordenadoria.

O problema consiste na classificação do texto da denúncia em uma das 8 coordenadorias existentes – posteriormente se verificou que uma denúncia pode ser classificada em mais de uma coordenadoria, hipótese em que pode haver desmembramento.

alias	descricao
01.	MEIO AMBIENTE DO TRABALHO
02.	TRABALHO ANÁLOGO AO DE ESCRAVO E TRÁFICO DE PESSOAS
03.	FRAUDES TRABALHISTAS
04.	TRABALHO NA ADMINISTRAÇÃO PÚBLICA
05.	TRABALHO PORTUÁRIO E AQUAVIÁRIO
06.	IGUALDADE DE OPORTUNIDADES, VIOLÊNCIA, ASSÉDIO E DISCRIMINAÇÃO NAS RELAÇÕES DE TRABALHO
07.	PROTEÇÃO DA CRIANÇA E DO ADOLESCENTE
08.	LIBERDADE E ORGANIZAÇÃO SINDICAL
09.	TEMAS GERAIS

Tabela 1: Coordenadorias (classes de assuntos). A classe 9 significa nenhum tema específico

A princípio parece se constituir em um problema de classificação em 9 classes, o que traz significativa complexidade. Porém, após análise, concluiu-se que se pode desmembrar o problema em 8 classificadores: cada um representando pertence, ou não pertence a uma coordenadoria, já que a denúncia pode conter mais de um assunto a ser investigado pelos procuradores (membros do MPT).

Foram feitas consultas nos dados, e foi constatado que o texto a ser classificado possui baixo tamanho (64kb no máximo, mas que em geral é menor que 10kb). Além disso, outros campos estruturados (categóricos) foram avaliados como significativos, como *possui_idosos_envolvidos*, *possui_criancas_envolvidas*, *criancas_numeroestimado*, *crianças_atividadesrealizadas*, *possui_deficientes_envolvidos*; pois podem influenciar na classificação. Neste sentido, pensou-se em adicionar alguma informação textual que traduza essas informações ao texto original da denúncia (no padrão 'possui idosos envolvidos: sim; possui crianças envolvidas:sim').

Conclusão: O problema constitui um desafio significativo e vale a pena ser estudado. Há dezenas de milhares de registros colhidos desde 2014 somente no estado do Rio Grande do Sul. **Uma barreira significativa é a possível existência de dados pessoais no corpo do texto da denúncia, o que acarreta cuidado no seu uso e na publicação do dataset para reprodutibilidade do trabalho** (ainda que tenham sidos excluídos os registros que foram marcados como tendo algum tipo de sigilo requerido pelo denunciante). Foi instaurado um pedido para utilização de um subconjunto dos dados para treinamento e teste ao Órgão por meio de um processo administrativo, mas seu deferimento depende de análise interna sem data limite para ocorrer.

Devido a problemática da anonimização, foi decidido trabalhar com dados textuais de enunciados de Peer Review na revisão de artigos no domínio científico (referência: https://www.researchgate.net/publication/370763519_PolitePEER_does_peer_review_hurt_A_dataset_to_gauge_politeness_intensity_in_the_peer_reviews)

2. Os dados escolhidos para o projeto prático em termos de suas principais características (tamanho, dimensionalidade, tipos de atributos, distribuição de instâncias por classe ou do atributo alvo numérico, distribuição dos atributos, etc.),

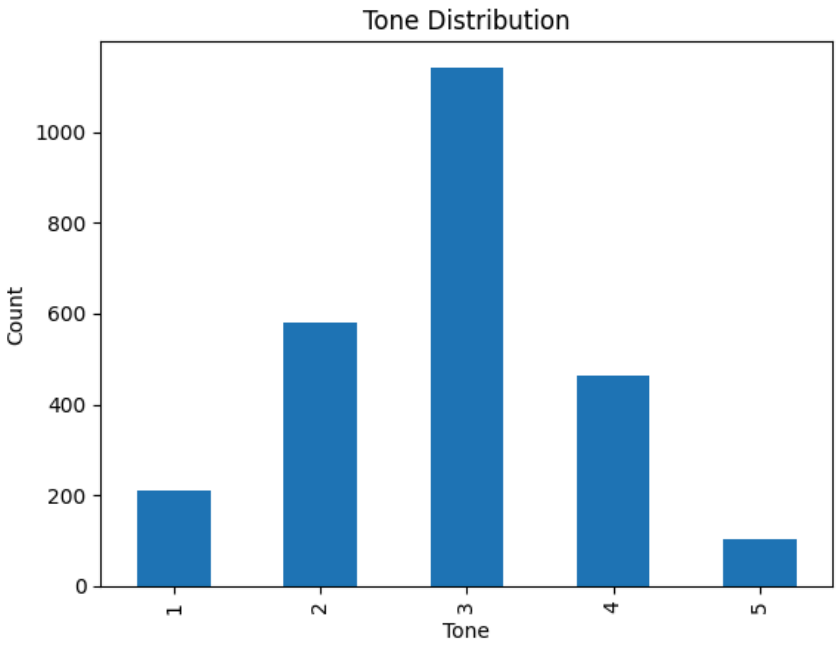
- A seguir é possível avaliar várias características sobre os dados empregados:

Tabela de tipo de dados e cardinalidade:

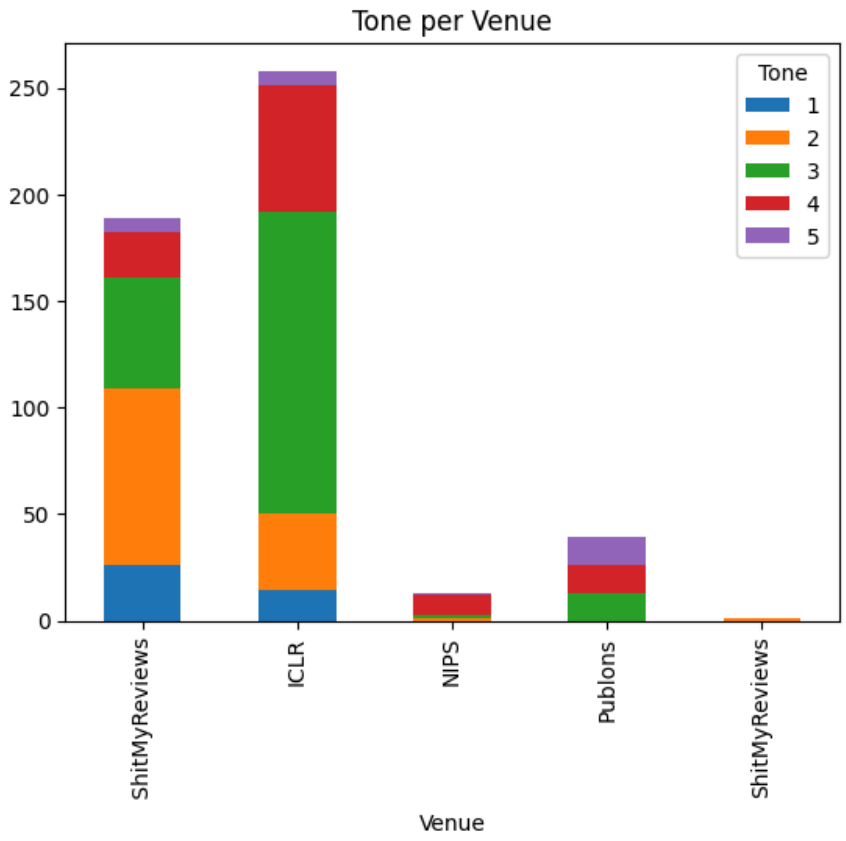
nome	cardinalidade	tipo
Venue	2500	categorizado
Review ID	1605	texto
review	2499	texto
Tone	2500	categorizado
Review URL	208	texto

Tabela da Classe alvo:

Tone	classe	count
1	highly impolite	210
2	impolite	581
3	neutral	1142
4	polite	465
5	highly polite	102



Distribuição de Tone (Classe Alvo)



Distribuição de Classes por Fonte de Reviews

3. Problemas já identificados nos dados através da análise exploratória

- Um problema identificado é a falta de outra features para treinar o modelo, talvez podem ser gerados novas features baseadas somente no enunciado da “review”

4. Estratégias que imaginam aplicar (ou já aplicaram) para tratar os problemas mencionados.

- Para pré processamento foi pensada a remoção de elementos não desejados para o treinamento, como pontuação e caracteres especiais
- Também foi-se transformado a coluna de “fonte” de reviews para numérico, pois os dados são desbalanceados por classe, o que faz sentido, porém para o treinamento precisaremos avaliar melhor este comportamento.
- A análise das embedding geradas também vai fazer parte do nosso estudo visto que vamos explorar alternativas não só como diferentes métodos de embeddings mas também diferentes modelos para avaliar a performance da classificação