Análise de Linguagem Ofensiva

Disciplina: Mineração de Textos e da Web, 2020.1

Equipe 02: Nicole Wirtzbiki, Agenor Júnior, Torricelli Evangelista

VISÃO GERAL DA ANÁLISE E CLASSIFICAÇÃO DOS TWEETS

1. Carregamento do Dataset de treino.
2. Carregamento dos Datasets de teste: testset_a e testset_b
3. Normalização dos Tweets (preprocessamento e limpeza dos dados)
4. Criação de Features Sintáticas (base de conhecimento internas)
5. Criação de Features Semânticas
    a. a partir de bases externas: Vetores GloVe e Empath.
6. Algoritmos de Machine Learning para Classificação dos Textos
7. Treino dos modelos para a subtask_a & Teste com o testset_a
8. Treino dos modelos para a subtask_b & Teste com o testset_b

FEATURES DA LINGUAGEM

Base de conhecimento interna:

- CONTAGEM DE PALAVRAS
- CONTAGEM DE CARACTERES
- COMPRIMENTO MÉDIO DAS PALAVRAS
- CONTAGEM DE STOPWORDS
- CONTAGEM DE #HASHTAGS
- CONTAGEM DE @MENTIONS
- CONTAGEM DE DÍGITOS NUMÉRICOS
- CONTAGEM DE PALAVRAS EM CAIXA ALTA

FEATURES SEMÂNTICAS

- BOW UNIGRAMA - VETORES DE FREQUÊNCIA
- TF-IDF - VETOR DE FREQUÊNCIA INTER-DOCUMENTOS

Word Embeddings - Vetores de Contexto de Uso a partir de Bases de Conhecimento Externas:

- Vetores GloVe (baseado em probabilidade, similaridade entre documentos)
- Vetores Empath (neural word embedding)

ALGORITMOS DE M.L. PARA A CLASSIFICAÇÃO DE TEXTO

- Naive Bayes
- Logistic Regression
- SVM - Support Vector Machine
- Random Forest

CROSS VALIDATION: Validação cruzada K-fold para balancear a flutuação dos resultados.

CONFUSION MATRIX:

		Detectada	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

FEATURES UTILIZADAS EM CADA SUBTASK

- BOW DE FREQUÊNCIA UNIGRAMA
- FEATURES DA LINGUAGEM
- BOW & FEATURES DA LINGUAGEM
- TF-IDF
- GLOVE
- GLOVE & EMPATH

APRESENTAÇÃO DO CÓDIGO NO NOTEBOOK