PUCRS - Escola Politécnica
Inteligência Artificial - Prof. Sílvia

Trabalho 2 - Análise de Sentimentos

1. **Objetivo do trabalho:** A análise de sentimentos é uma subárea do processamento de linguagem natural que visa classificar as opiniões de pessoas sobre produtos, serviços, marcas, etc. Neste tipo de abordagem os textos que expressões opiniões são classificados em positivo, negativo ou neutro. Os textos classificados com positivos expressam uma opinião favorável; os negativos, desfavorável; e os neutros não expressam opinião. O objetivo deste trabalho é estudar análise de sentimentos, desenvolvendo um experimento sobre o corpus extraído do Twitter.

2. **Corpus**: é um nome dado a uma coleção de documentos (textos, sentenças, etc). Neste trabalho será usado o corpus COMPUTER-BR (https://github.com/silviawmoraes/PLN/tree/master/corpora/computer-br).

3. **Etapas do trabalho** - Abaixo as macro-etapas do trabalho:

   (a) Pré-processamento do corpus: normalização morfológica, anotação linguística, extração dos termos, seleção dos termos mais relevantes, estruturação.

   (b) Categorização e análise dos resultados.

   (c) Escrita de um relatório sobre o trabalho realizado.

4. **Descrição da Etapa de Pré-processamento:** O pré-processamento é a etapa mais custosa de qualquer tarefa em Aprendizagem de Máquina (AM). A preparação dos textos é ainda um pouco mais custosa, pois textos são dados desestruturados. Para estruturá-los, ou seja, colocá-los em um formato que viabilize o processamento dos mesmos por um algoritmo de AM:

   (a) Normalização Morfológica e Anotação Linguística: Pode ser feita pelo parser VISL, Cogroo, Tree Tagger, nltk. O objetivo da normalização morfológica é colocar os termos (strings) na mesma "forma". Por exemplo, verbos quando aparecem nos textos estão flexionados: "estudou", "estudaram" e "estuda". A meta, nesse caso, é transformar essas ocorrências em uma forma normal, que no caso dos verbos é o infinito: "estudar". Existem vários tipos de normalização morfológica. As mais comuns são lematização e stemming. O objetivo da anotação linguistica é prover informações sobre o texto para que possamos, em um segundo momento, escolher os termos mais relevantes de um texto. Anotar um texto é colocar tags (rótulos) em seus termos. Essas tags podem ser morfossintáticas e, até mesmo, semânticas. Nesse trabalho, vamos usar apenas a anotação de Part-Of-Speech (POS). As tags de POS indicam as classes gramaticais das palavras: verbo (V), substantivo (N, PROP) adjetivo (ADJ) e advérbio (ADV).

   (b) Extração do Termos: Após o processo de anotação, já podemos retirar do texto termos que podem ser úteis na etapa de estruturação. Para cada classe do corpus, construa uma lista com os tokens mais relevantes. Preserve, em sua implementação, a informação sobre o texto do qual esses termos foram extraídos. Faz parte do seu trabalho identificar as classes gramaticais mais relevantes para este trabalho.

   (c) Seleção dos Termos mais relevantes: É nessa etapa que precisamos escolher os termos mais relevantes (redução de dimensionalidade) visando a representação dos textos (estruturação). Usando as listas criadas na etapa anterior, crie uma lista geral de termos (sem repetição). Para cada termo dessa lista, contabilize a frequência desse termo no corpus. A seguir, selecione os *k* primeiros termos mais frequentes. Faz parte do seu trabalho definir o valor de *k* mais adequado. O resultado dessa seleção é uma lista de termos, conhecida como Bag-of-Words (BoW).

(d) Estruturação: Nessa fase, vamos usar uma representação vetorial para estruturar os textos. A BoW funcionará com os atributos (campos) do texto. A representação vetorial mais simples é a binária, que indica se um termo da BoW está ou não no texto. Por exemplo, supondo que a BoW é formada pelo vetor [P1,P2,P3,P4] e existem os seguintes textos já pré-processados (cada um com sua lista de termos): T1 = {P4, P5, P6}; T2 = {P1,P3, P7}; T3 = {P8,P4, P5} ; T4 = {P1,P8, P9}e T5 = {P1,P4, P9} . Considerando que T1,T2 e T3 são Positivos e T4 e T5, negativos. Os textos estruturados, ficaria assim:

P1 P2 P3 P4

0, 0, 0, 1, Positivo (T1)
1, 0, 1, 0, Positivo (T2)
0, 0, 0, 1, Positivo
1, 0, 0, 0, Negativo
1, 0, 0, 1, Negativo

5. **Descrição da Etapa de Classificação:** A categorização (ou classificação) de textos é o processo de automaticamente atribuir uma ou mais categorias predefinidas a documentos textuais. Nessa etapa testaremos algoritmos de classificação sobre o corpus pré-processado. O objetivo dessa etapa é testar com diferentes algoritmos (ao menos 3, incluindo k-NN) e comentar aquele que melhor classificou os textos. Usar, nessa etapa, 80% dos textos (de cada classe) para treino e os restantes para teste. Nos 80% utilize validação cruzada para definir o modelo.

6. **Descrição do Relatório:** Deve ser entregue um relatório descrevendo: o objetivo do trabalho, pré-processamento (descrever o pré-processamento realizado, configurações da BoW); para cada tarefa, mencionar os algoritmos testados e detalhar a análise dos resulta dos (tomar como base as medidas usuais), bem como incluir comentários sobre o desenvolvimento do mesmo e a sua conclusão.

7. **Desenvolvimento e Entrega:** O trabalho poderá ser desenvolvido ao longo das aulas práticas da disciplina. Entrega final: 23/06/2022

8. **Forma de avaliação**
   - Etapas de pre-processamento (da normalização à estruturação): 3,0 pontos
   - Etapa de Aprendizagem - tarefa de Categorização : 5,0 pontos
   - Análise dos resultados (relatório): 2,0 pts