Mineração de Opiniões

Prof. Dr. Tiago Eugenio de Melo (EST/UEA)
tmelo@uea.edu.br

Aula
Sentenças Comparativas

Referências Bibliográficas

- KANSAON, Daniel et al. **Mining Portuguese Comparative Sentences in Online Reviews**. In: Proceedings of the Brazilian Symposium on Multimedia and the Web. 2020. p. 333-340.
- LIU, Bing. **Sentiment analysis and opinion mining**. Synthesis lectures on human language technologies, v. 5, n. 1, p. 1-167, 2012.

Objetivos

- Apresentar a descrição do problema de identificação de sentenças comparativas
- Apresentar um trabalho prático

Introdução

- O número de compras no mercado online vem aumentando e estima-se que cerca de 25% da população mundial utilizará esse mercado nos próximos anos.
- As opiniões podem ser:
  - Regulares
  - Comparativas
- Uma grande parte dos trabalhos na literatura se concentram na aplicação de técnicas de análise de sentimentos para classificação de sentenças em positivas, negativas e neutras.

Conceitos

As comparações podem ser separadas em dois principais grupos:

- **Comparações gradativas** que expressam relação de ordem entre as entidades comparadas nas sentenças, podendo ser de semelhança ou de superioridade.
- **Comparações não gradativas** que comparam objetos sem expressar ordem entre eles.

Conceitos

- Bing Liu define quatro categorias para organizar as opiniões comparativas, as três primeiras fazem parte das comparações gradativas, já a última das comparações não gradativas.
- Categorias:
  - Gradativa com Predileção
  - Equitativa
  - Superlativa
  - Não Gradativa

Gradativa com Predileção

- Contém ao menos duas entidades expressando **predileção** e **ordem** de uma em relação à outra.
- Exemplo:
  - *O carro X é melhor que o carro Y*.

Equitativa

- Existem duas entidades na qual a relação entre elas é de **igualdade** baseada em algum aspecto.
- Exemplo:
  - *A câmera do smartphone X é igual ao Y.*

Superlativa

- Uma entidade possui relações do tipo **maior ou menor** que um grupo de outras.
- Exemplo:
  - *Este é o melhor laptop do mundo*.

Não Gradativa

- Compara duas ou mais entidades, mas **não expressa ordem** nem **predileção** por nenhuma.
- Exemplo:
  - *O design do laptop X possui alguns recursos diferentes do laptop Y.*

Metodologia proposta por Kansaon et al. [1]

- Construção de um léxico em português.
- Construção da base de dados comparativas.
- Etapa de classificação.

Construção do Léxico

- A hipótese é que existe um **grupo restrito de palavras e expressões** que é constantemente utilizado para fazer comparações.
- Foi construído um léxico disponível em [1].

[1] https://zenodo.org/record/4124410#.YNJqh217kTs

Construção do Léxico

- Palavras-chaves mais frequentes:

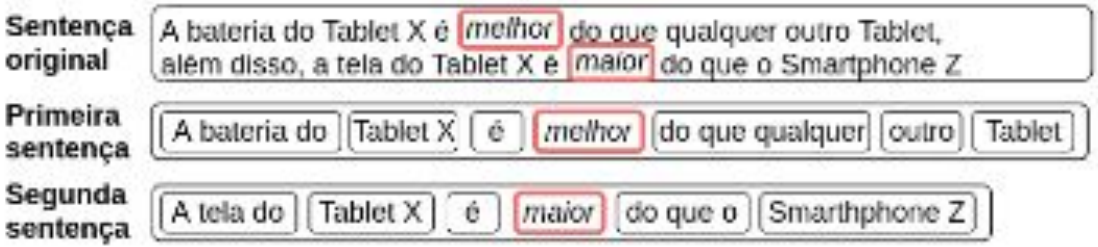
Dados	Palavras-Chave				
Buscapé	mais	como	recomendo	melhor	comprei
Twitter	mais	como	queria	melhor	parece

- Palavras-chaves mais precisas:

Dados	Palavras-Chave				
Buscapé	incomparável	líder	idêntico	assemelha	supera
Twitter	incomparável	similar	idênticos	assemelha	preferível

Construção da Base de Dados Comparativa

- Estratégia:

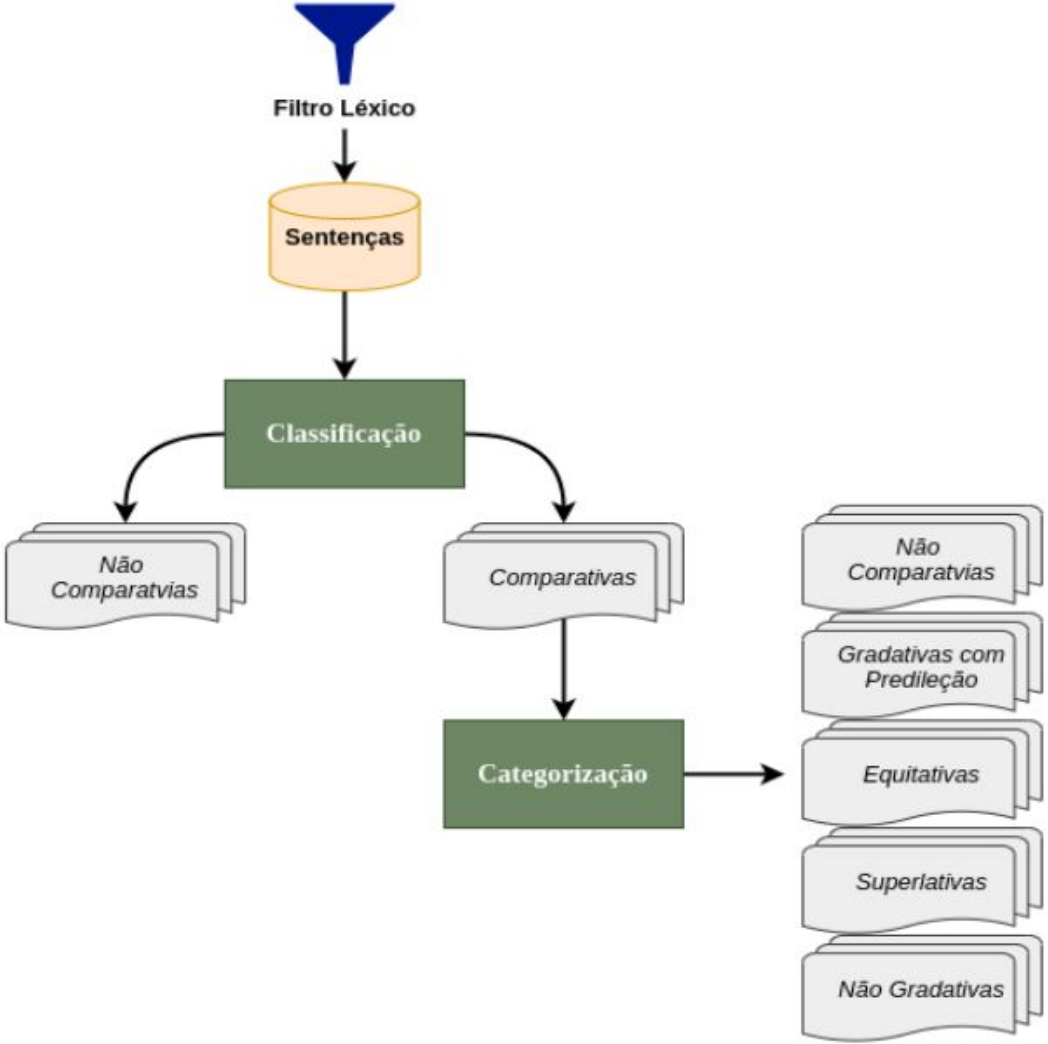


Construção da Base de Dados Comparativa

- Dataset:

Sentenças	Buscapé	Twitter	Total
Comparativas	1.282	918	2.200
Não Comparativas	1.472	1.135	2.607
Total	2.754	2.053	4.807

Problema


Filtro Léxico
Sentenças
Classificação
Não Comparatvias
Comparativas
Categorização
Não Comparatvias
Gradativas com Predileção
Equitativas
Superlativas
Não Gradativas

Problema

- Escolher um método supervisionado;
- Usar o mesmo protocolo experimental do artigo base;
- Desenvolver e avaliar um classificador:
  - Identificação de sentenças comparativas;
  - Classificação de múltiplas classes;

Dataset Problema

Sentenças	Buscapé	Twitter	Total
Gradativa com Predileção	502	279	781
Equitativa	255	172	427
Superlativa	290	270	560
Não Gradativa	115	81	196
Total Comparativas	1.162	802	1.964
Total Não Comparativas	234	170	404

Baseline

- Tarefa 1 (Buscapé):

	Buscapé							
	Não Comparativa			Comparativa			Média	
	Prec.	Rec.	F1	Prec.	Rec.	F1	ACC.	Macro F1
RF	0.741 ± 0.006	0.801 ± 0.008	0.77 ± 0.006	0.749 ± 0.008	0.679 ± 0.01	0.712 ± 0.007	0.744 ± 0.006	0.741 ± 0.006
LR	0.861 ± 0.006	0.863 ± 0.007	0.862 ± 0.005	0.843 ± 0.007	0.839 ± 0.008	0.841 ± 0.006	0.852 ± 0.005	0.851 ± 0.006
SVM	0.869 ± 0.005	**0.895** ± 0.006	**0.882** ± 0.004	**0.875** ± 0.007	0.845 ± 0.006	**0.86** ± 0.005	**0.872** ± 0.005	**0.871** ± 0.005
NB	**0.909** ± 0.005	0.847 ± 0.006	**0.877** ± 0.005	0.838 ± 0.006	**0.903** ± 0.006	**0.869** ± 0.005	**0.873** ± 0.004	**0.873** ± 0.004

Baseline

- Tarefa 1 (Twitter):

	Twitter							
	Não Comparativa			Comparativa			Média	
	Prec.	Rec.	F1	Prec.	Rec.	F1	ACC.	Macro F1
RF	0.741 ± 0.007	0.84 ± 0.011	0.787 ± 0.008	0.764 ± 0.013	0.637 ± 0.011	0.695 ± 0.01	0.749 ± 0.008	0.741 ± 0.009
LR	0.831 ± 0.006	0.874 ± 0.007	0.851 ± 0.005	0.833 ± 0.008	0.779 ± 0.01	0.805 ± 0.007	0.831 ± 0.006	0.828 ± 0.006
SVM	0.834 ± 0.007	**0.912** ± 0.005	**0.871** ± 0.005	**0.878** ± 0.007	0.775 ± 0.011	0.823 ± 0.007	**0.851** ± 0.006	0.847 ± 0.006
NB	**0.894** ± 0.007	0.851 ± 0.008	**0.872** ± 0.005	0.827 ± 0.007	**0.874** ± 0.009	**0.85** ± 0.006	**0.862** ± 0.005	**0.861** ± 0.005

Baseline

- Tarefa 2 (Buscapé):

Buscapé						
		Label Predito				
		Não Comparativa	Grad. com Pred.	Equitativa	Superlativa	Não Gradativa
Label Real	Não Comparativa	**34,3%**	29,7%	13,1%	14,2%	8,7%
	Gradativa com Predileção	9,8%	**75,1%**	5,4%	5,7%	4,0%
	Equitativa	11,3%	9,9%	**74,3%**	2,3%	2,2%
	Superlativa	5,9%	11,6%	2,3%	**79,8%**	0,5%
	Não Gradativa	17,3%	22,7%	10,6%	3,7%	**45,7%**

Baseline

- Tarefa 2 (Twitter):

Twitter

		Label Predito				
		Não Comparativa	Grad. com Pred.	Equitativa	Superlativa	Não Gradativa
Label Real	Não Comparativa	**37,9%**	21%	15,9%	20,7%	4,5%
	Gradativa com Predileção	5,6%	**85,6%**	3,7%	4,9%	0,3%
	Equitativa	11,0%	7,8%	**78,9%**	2,2%	0,1%
	Superlativa	13,8%	16,5%	2,5%	**66,9%**	0,3%
	Não Gradativa	33,0%	0,7%	30,2%	0,5%	**35,6%**

Regras

- Escolha de um único algoritmo.
- Apresentação das estratégias adotadas.
- Apresentação dos resultados alcançados.
- Comparação com *baseline*.
- Trabalho em duplas.
- Dataset: https://drive.google.com/file/d/1gGnHMAqoFAt4Mi-iyORNPPNHEV6KPMdy/view?usp=sharing
- Dia 28/06 -> tirar dúvidas.
- Dia 29/06 às 23:59 -> limite para submissão do trabalho.
- Dia 30/06 -> apresentação do trabalho.