Toolbox de processamento de linguagem natural com *scikit-learn*

Comparativo de algoritmos e técnicas de pré-processamento

Agenda

- Definições
- Bases de dados, modelos e pré-processamento
- Experimentos e resultados
- Conclusões

Definições

- Aprendizado supervisionado com *Decision Tree*, *Naive Bayes* e *Support-Vector Machine*

- *Count Vectorizer*

- tf-idf

- *N-grams*

- *Stopwords*

- *F1 score*

Bases de dados

- Análise de sentimentos de revisões (*reviews*) no IMDb
  Classes: “positivo” ou “negativo”
  Total de 49.459 objetos
- Classificação de produtos com base em sua descrição
  Classes: “livro”, “game”, “maquiagem” e “brinquedo”
  Total de 2916 objetos

Modelos utilizados

- *Decision Tree*
- *Naive Bayes*
- *Support-Vector Machine* (SVM)

Pré-processamento

- *Count Vectorizer*

- tf-idf

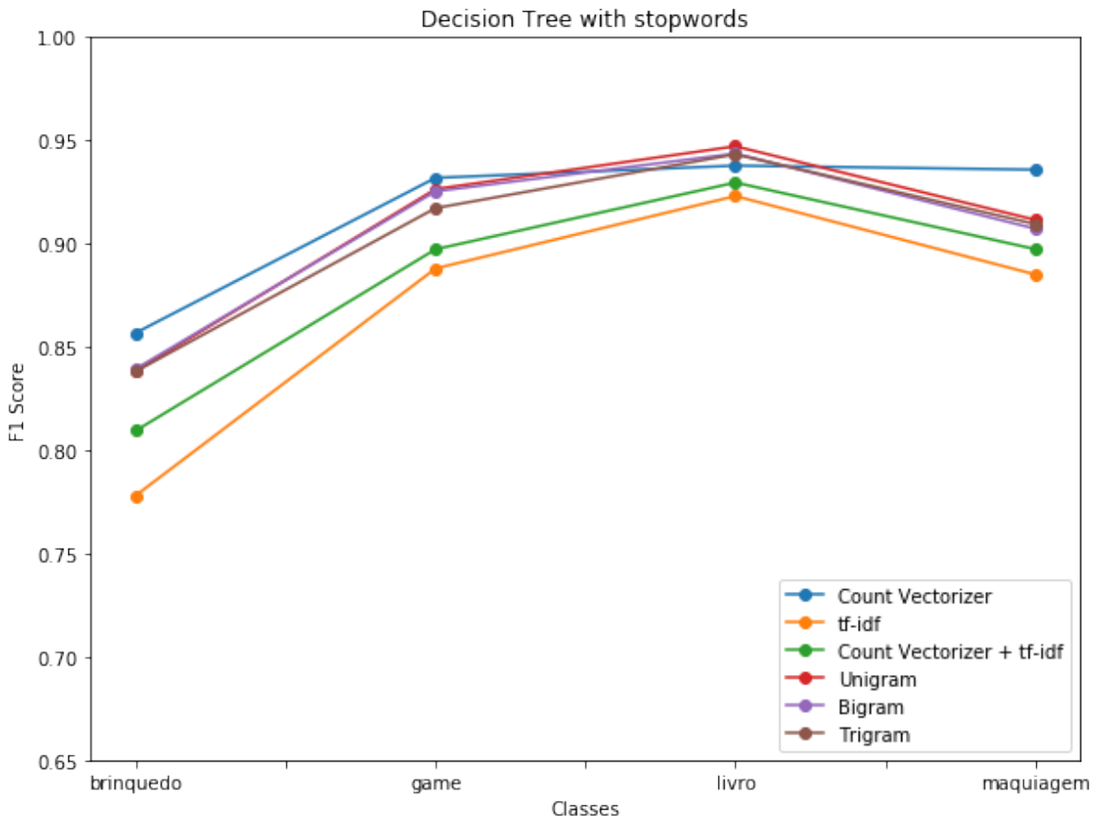
- *Count Vectorizer* + tf-idf

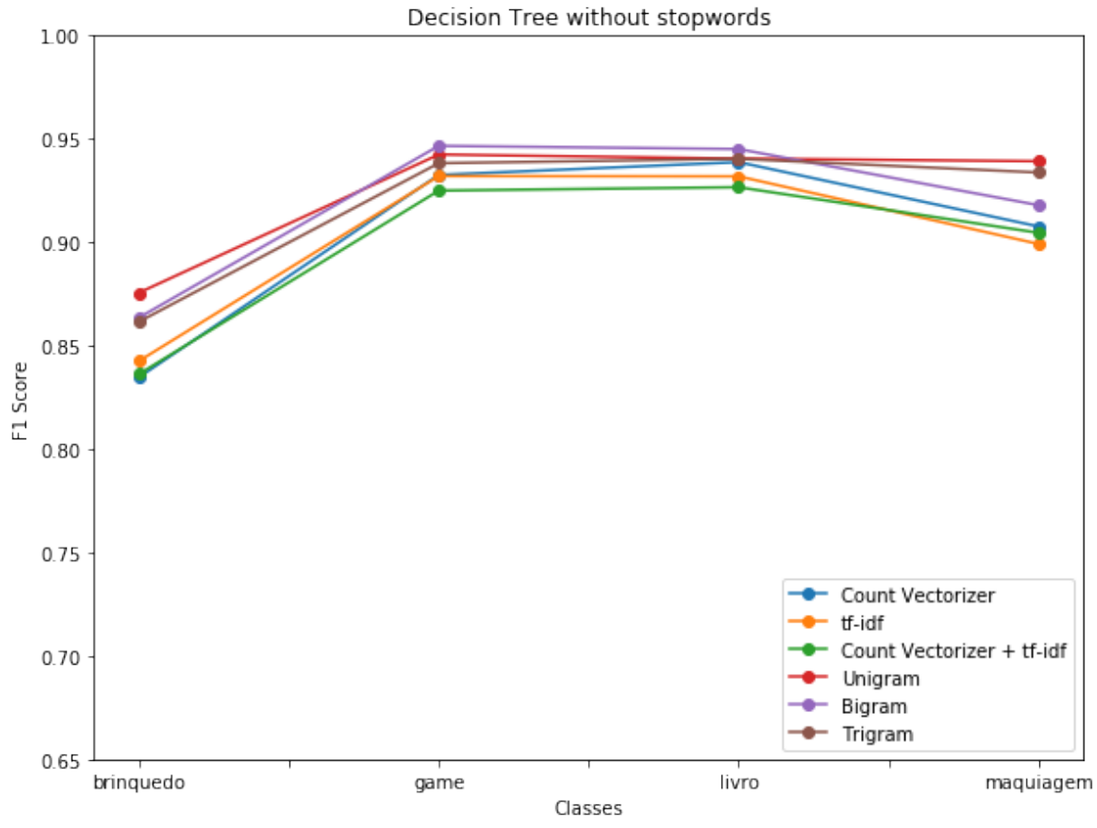
- *N-grams*:
  1-gram (unigram),
  2-grams (bigram) e
  3-grams (trigram)

Experimentos

- 2 bases de dados
- 6 x 2 (12) métodos de pré-processamento
  Com *stopwords* e sem *stopwords*
- 3 modelos de aprendizado supervisionado
- Total de 72 experimentos
- Acurácia a partir do *F1 score*

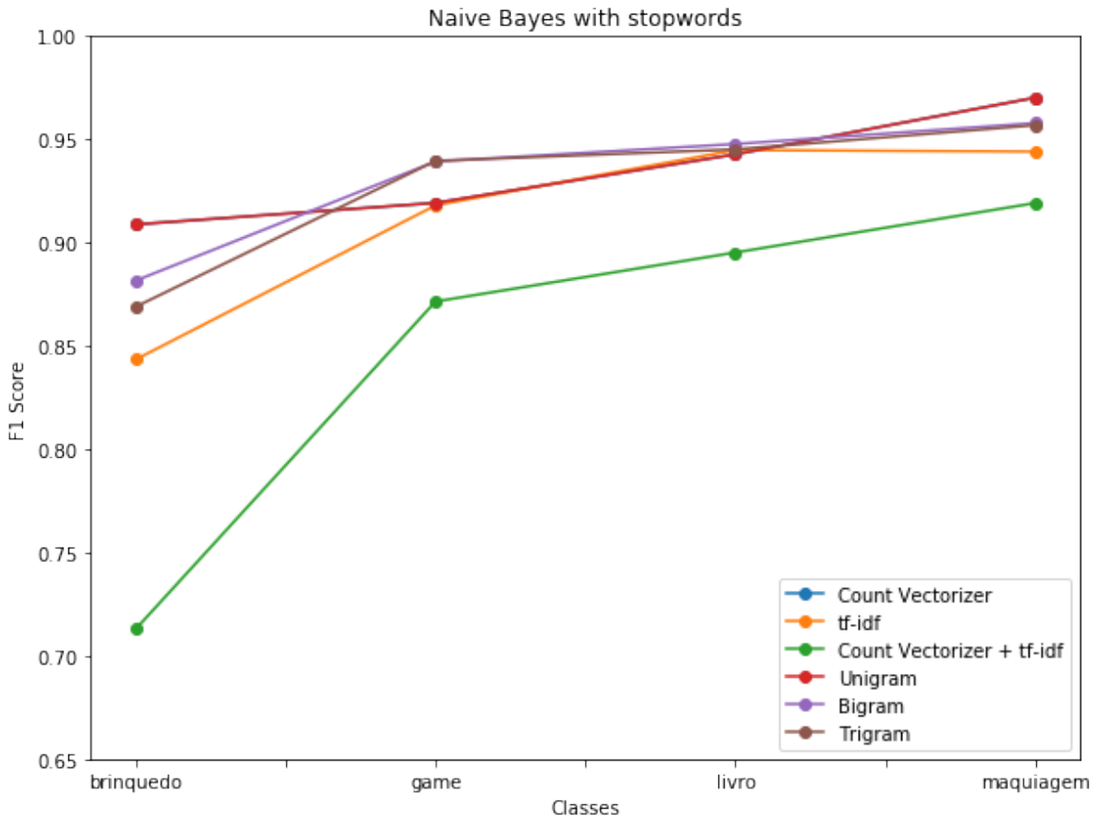
Resultados

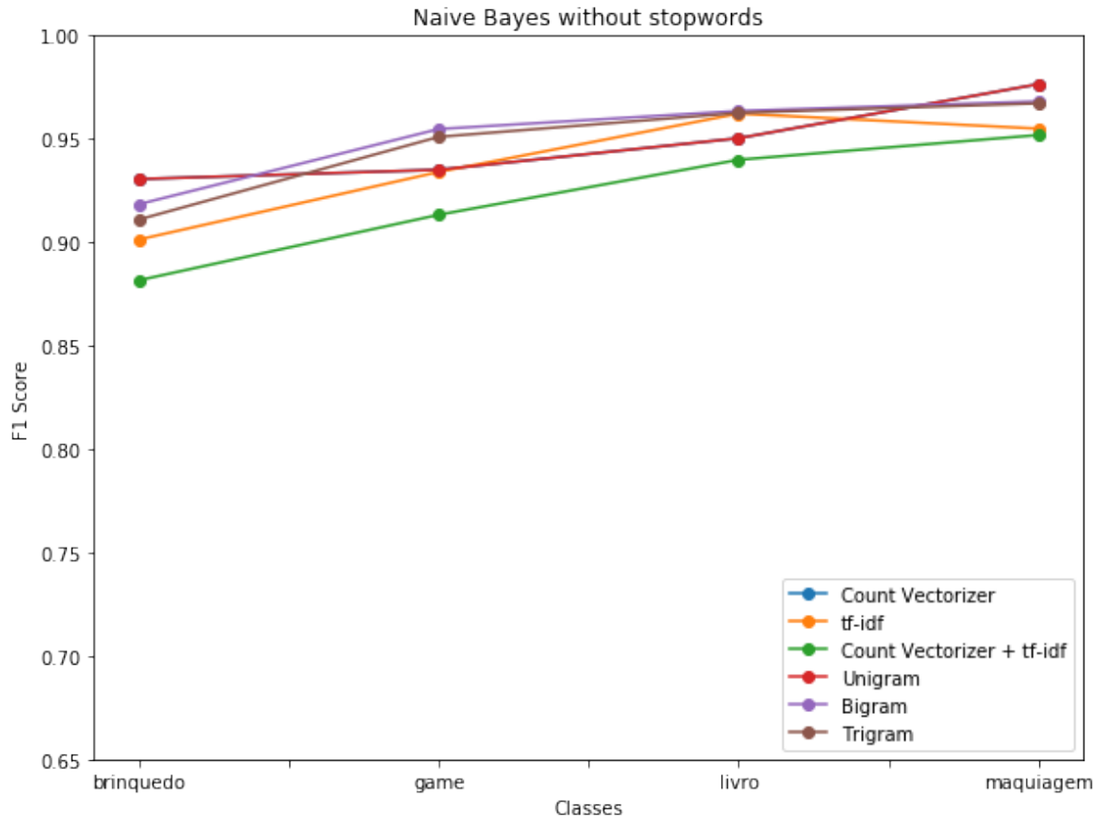




Classificação de produtos com *Decision Tree*,
com presença das *stopwords* (esquerda) e com remoção destas (direita)

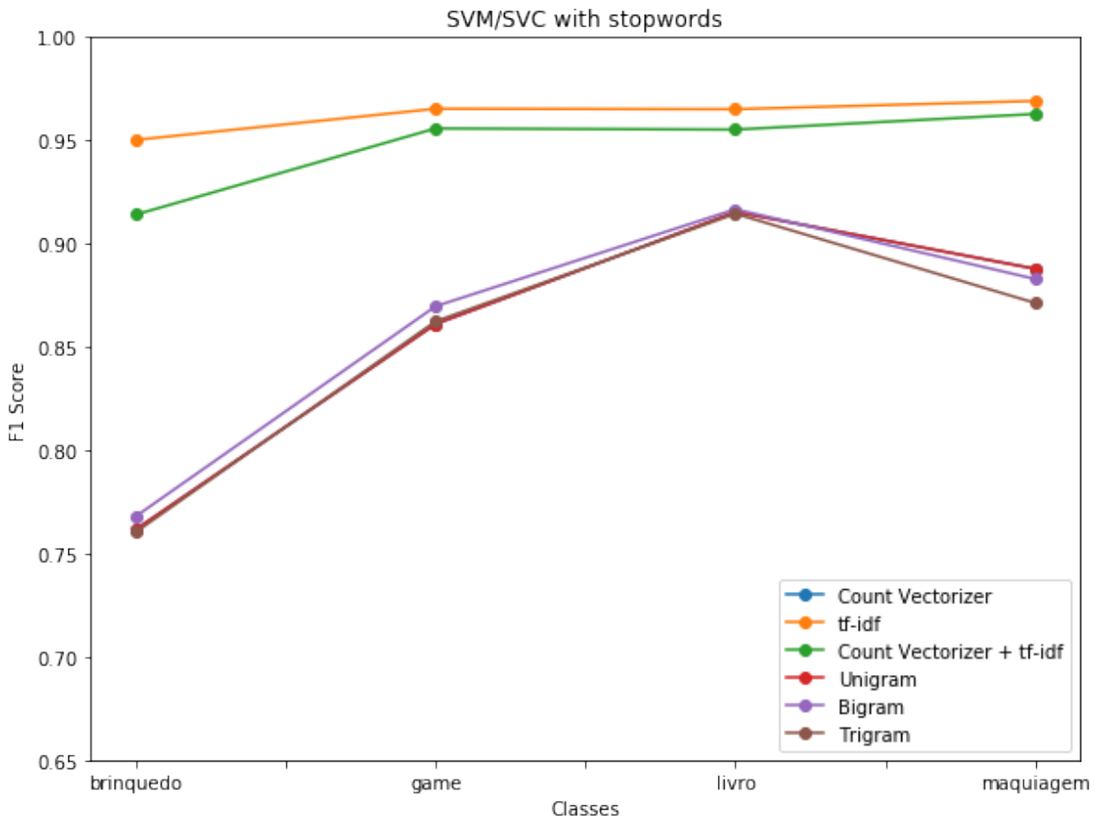
Resultados

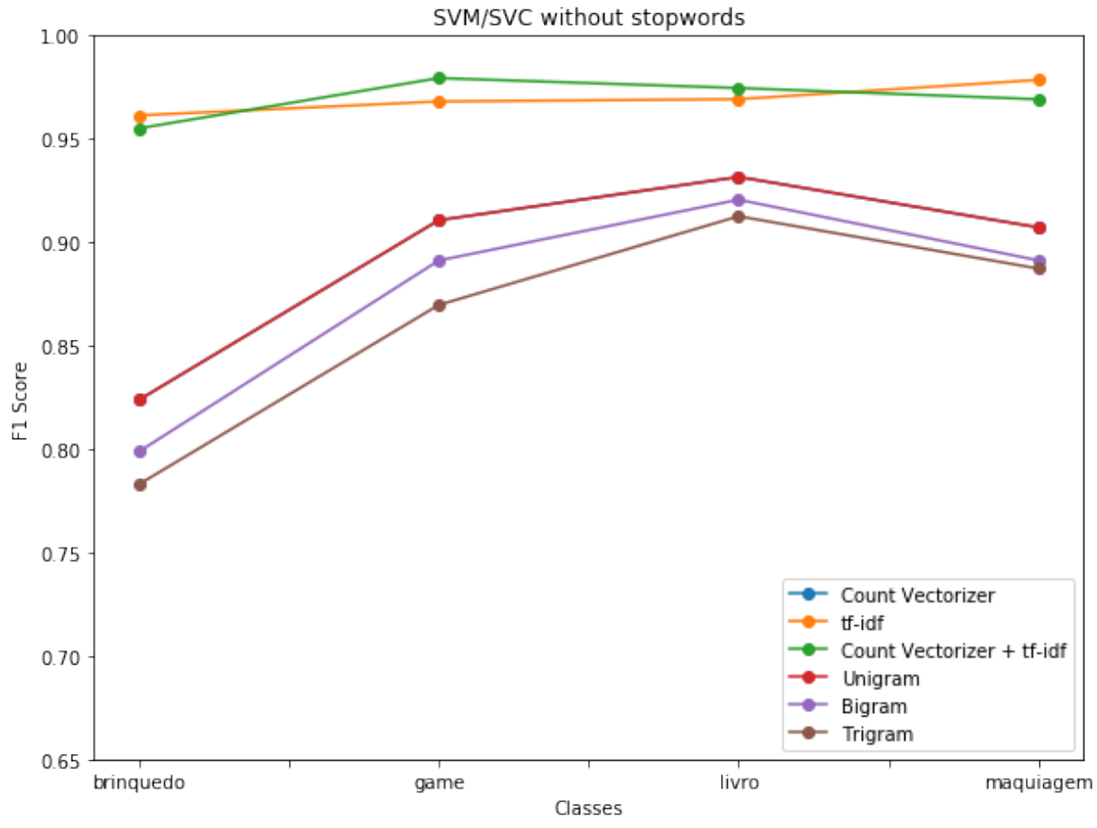




Classificação de produtos com *Naive Bayes*,
com presença das *stopwords* (esquerda) e com remoção destas (direita)

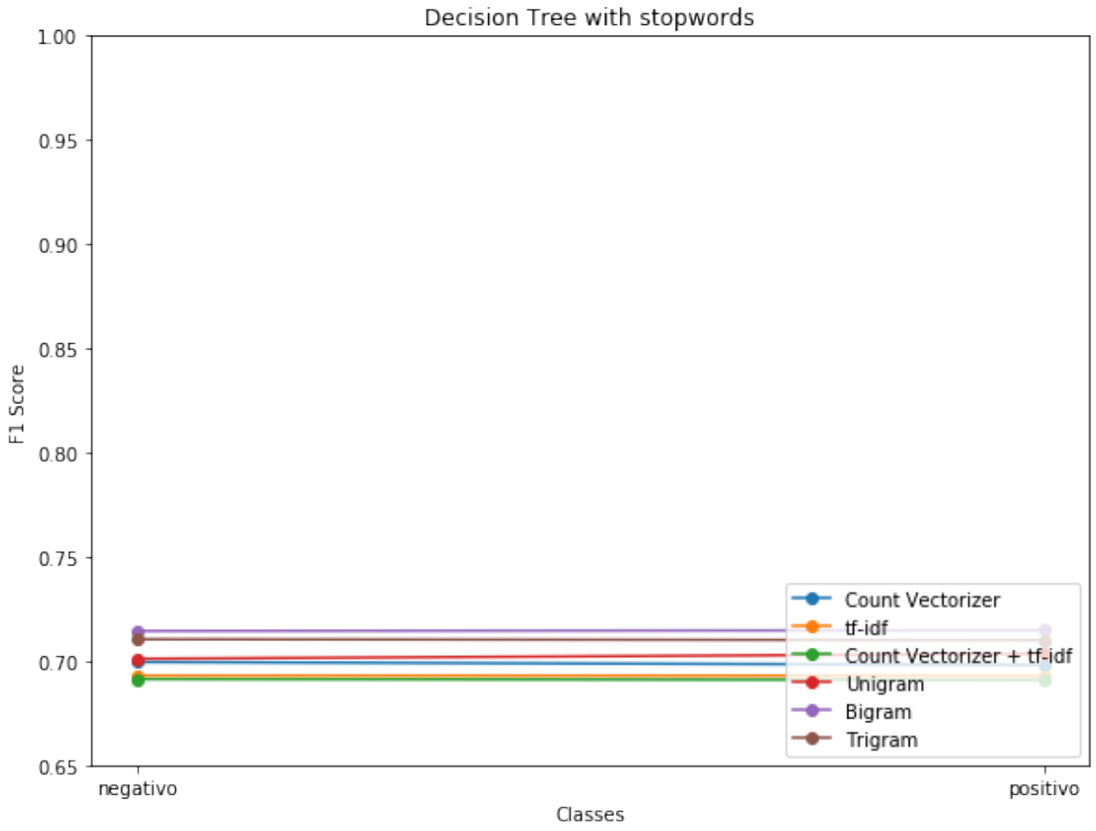
Resultados

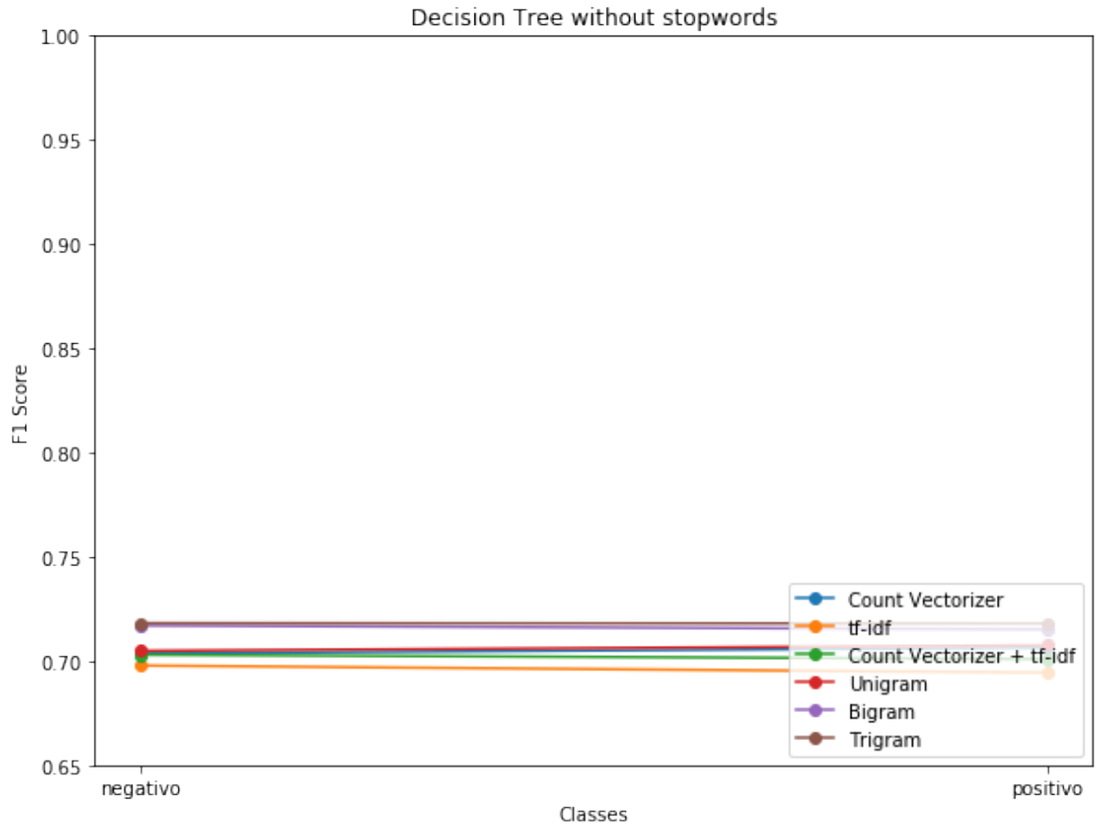




Classificação de produtos com *Support-Vector Machine*,
com presença das *stopwords* (esquerda) e com remoção destas (direita)

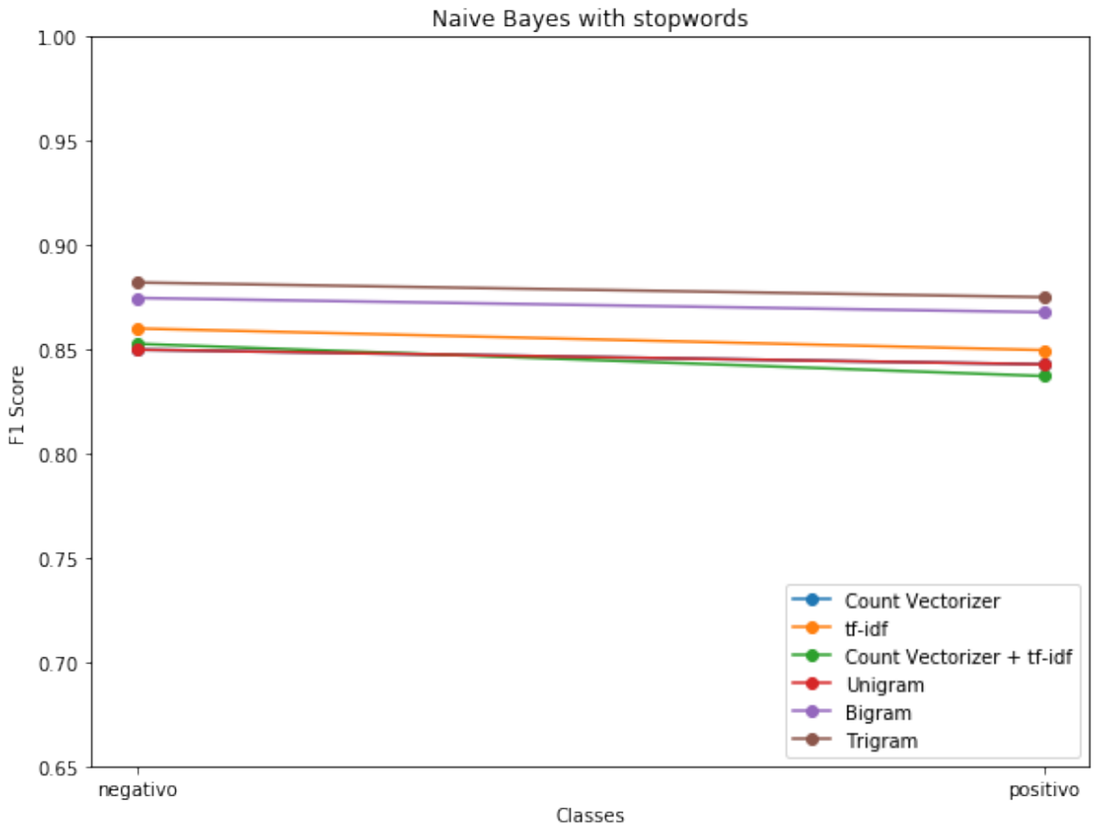
Resultados

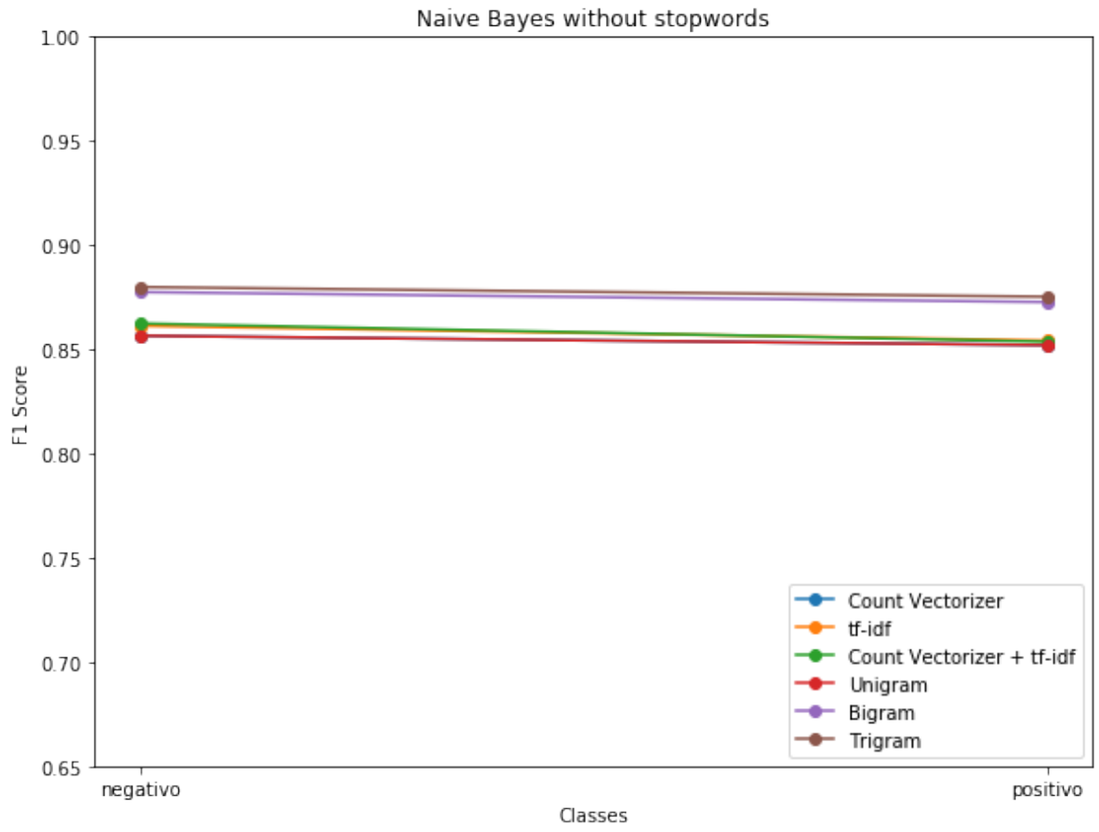




Análise de sentimento com *Decision Tree*,
com presença das *stopwords* (esquerda) e com remoção destas (direita)

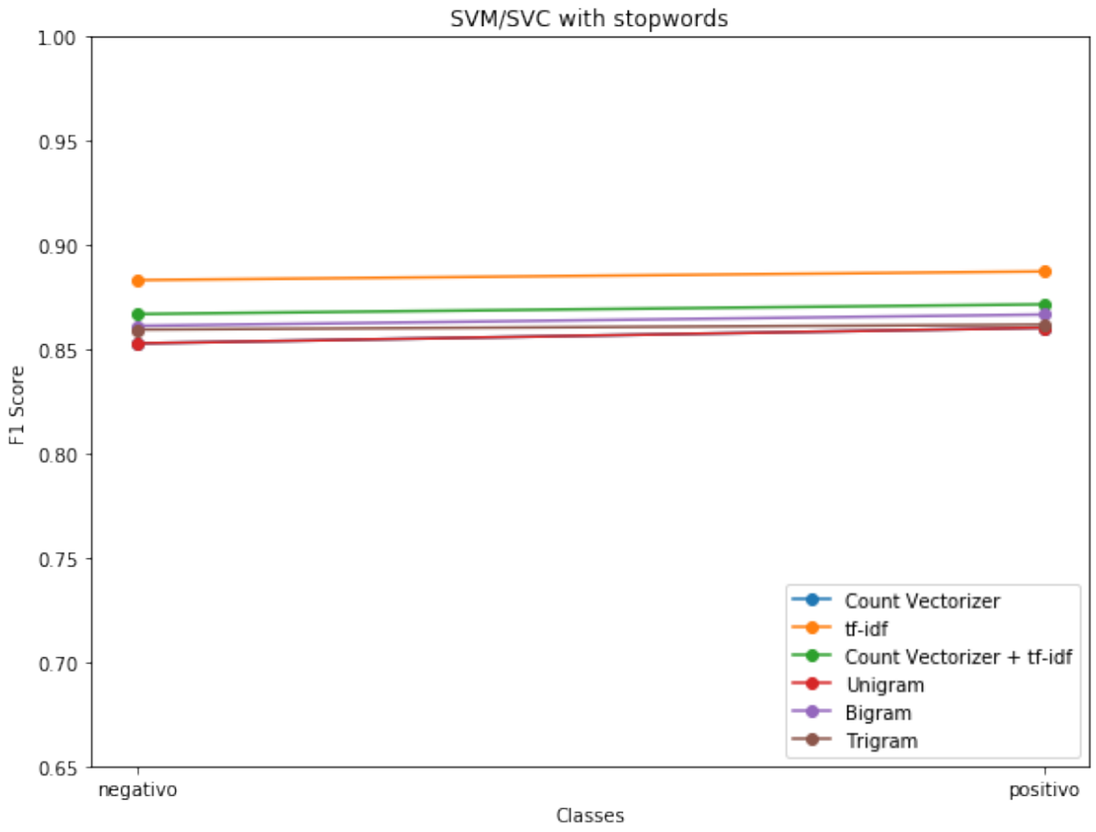
Resultados

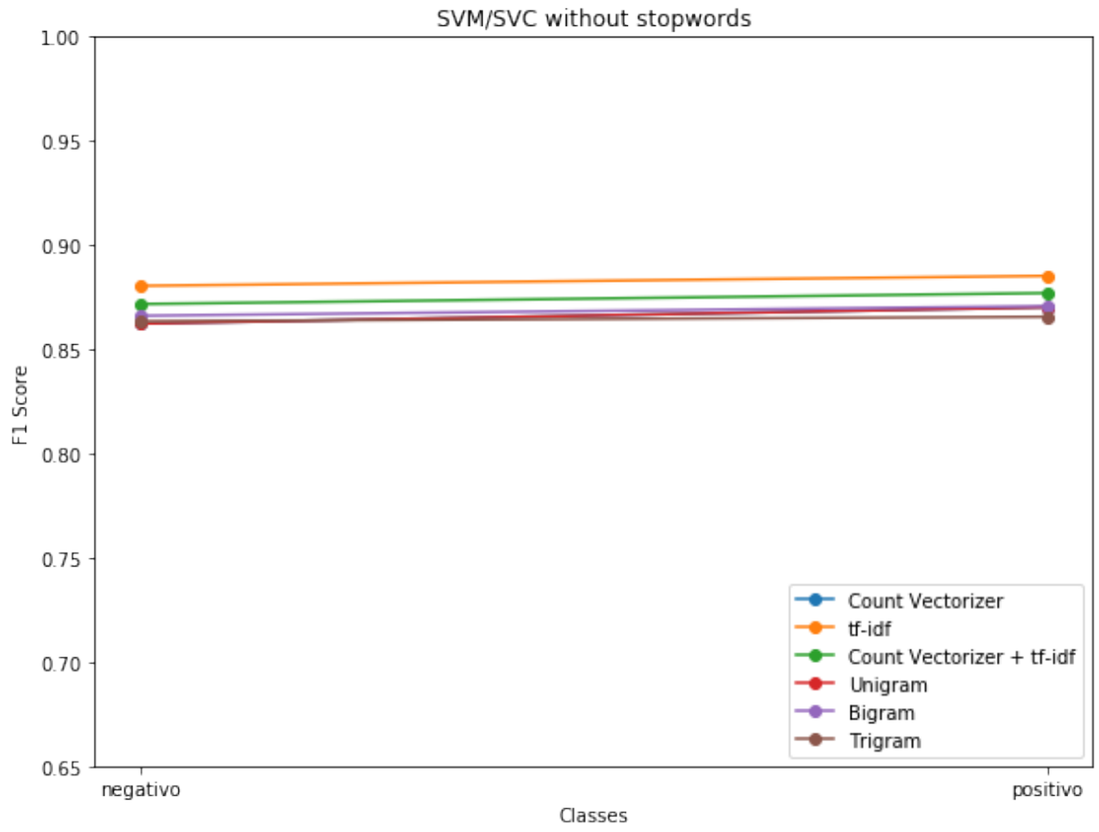




Análise de sentimento com *Naive Bayes*,
com presença das *stopwords* (esquerda) e com remoção destas (direita)

Resultados





Análise de sentimento com *Support-Vector Machine*,
com presença das *stopwords* (esquerda) e com remoção destas (direita)

Conclusões

- Em geral, resultados de qualidade com SVM + tf-idf
- Variação na qualidade dos resultados com a remoção das *stopwords:* depende do modelo, do tipo de pré-processamento e das classes do problema
- Relevância da análise da base de dados e das suas características próprias ao escolher o tipo de pré-processamento e modelo de aprendizado

Obrigado!

Ewerton Carlos Assis
carlos.assis@ufabc.edu.br