Análise de Sentimentos usando Machine Learning

Gabriel Radzki França[1]

[1]Escola Politécnica – Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Av. Ipiranga, Prédio 30 - Bloco F - Partenon – 90.619-900 – Porto Alegre – RS – Brazil

gabriel.radzki@edu.pucrs.br

1. Implementação

Para esse trabalho, desenvolvi um programa em Python que faz o pré-processamento e a estruturação para uso dos dados no Weka. A primeira parte é, de fato, a mais trabalhosa. O fato do *corpus* ser composto por tweets, faz com que a linguagem seja quase imprevisível, e, por isso, necessita de um cuidado maior na hora de lidarmos com o *corpus*.

O programa divide o dataset em dois: **opiniões negativas** e **opiniões positivas**. Após isso, todas as *strings* são convertidas para caixa baixa. Em seguida, todos os *hyperlinks* são removidos, juntamente com *hashtags* e menções de usuário, e.g., "@pucrs".

Inicialmente, todos os acentos das strings foram removidos. O problema é que isso implica na perda de morfologia de algumas palavras, como o verbo **ser** (no tempo presente: "**é**"), que passa a ser uma conjunção. Portanto, os acentos (assim como o cedilha) foram mantidos em todas as palavras.

Letras repetidas em sequência (2+) são reduzidas para somente duas vezes, em uma tentativa de mitigar palavras com muita ênfase, e.g., "muitooooooo", ou "lixooooooo". A redução, transforma-as para "muitoo" e "lixoo", o que não é ortograficamente correto, mas supondo que existam duas versões da mesma palavra, i.e., "muitooooooo" e "muitooo", após a conversão elas terão a mesma grafia.

Algumas palavras foram corrigidas manualmente, após análise da distribuição de frequência das K *strings* mais utilizadas. Um exemplo de correção é a flexão informal "not", que se refere a palavra "notebook". Todas as ocorrências de "not" foram substituídas.

Após o tratamento das palavras, a normalização segue a seguinte sequência:

1) Extração das *Parts-Of-Speech* o software *TreeTagger* foi utilizado, junto ao *dataset* em português;
2) Tokenização com a biblioteca NLTK;
3) *Stemming* RSLP, que é uma versão em português, para o projeto NLTK, também.

Finalmente, é gerado o arquivo de saída (entrada, para o Weka), usando *Bag-Of-Words* de tweets positivos e negativos.

Algumas modificações foram feitas para testar algumas teorias:

Teoria 1: Hashtags são úteis para identificar sentimentos negativos
Analisando o *corpus*, percebi (empiricamente) que hashtags apareciam mais frequentemente em tweets negativos. Para testar, removi o caractere sustenido e mantive as palavras das hashtags. O resultado, ainda assim, foi uma porcentagem menor do que a explicitada nos resultados.

1.1 Estrutura de Dados

Os dados foram estruturados conforme o enunciado do trabalho:

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
```

2. Resultados

Teste 1 - Parâmetros:

a) SEM MERGE NOT ⇔ NOTEBOOK
b) COM HASHTAGS
c) K = **5**

```
=== Summary ===

Correctly Classified Instances          76               59.375  %
Incorrectly Classified Instances        52               40.625  %
Kappa statistic                          0.011
Mean absolute error                      0.432
Root mean squared error                  0.4901
Relative absolute error                 97.0294 %
Root relative squared error            100.509  %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,170    0,160    0,381      0,170    0,235      0,013    0,580     0,421     POSITIVE
                 0,840    0,830    0,636      0,840    0,723      0,013    0,580     0,698     NEGATIVE
Weighted Avg.    0,594    0,584    0,542      0,594    0,544      0,013    0,580     0,596

=== Confusion Matrix ===

  a  b   <-- classified as
  8 39 |  a = POSITIVE
 13 68 |  b = NEGATIVE
```

KNN

```
=== Summary ===

Correctly Classified Instances          79               61.7188 %
Incorrectly Classified Instances        49               38.2813 %
Kappa statistic                          0.0339
Mean absolute error                      0.4366
Root mean squared error                  0.4803
Relative absolute error                 98.0631 %
Root relative squared error             98.513  %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,128    0,099    0,429      0,128    0,197      0,045    0,580     0,425     POSITIVE
                 0,901    0,872    0,640      0,901    0,749      0,045    0,580     0,711     NEGATIVE
Weighted Avg.    0,617    0,588    0,563      0,617    0,546      0,045    0,580     0,606

=== Confusion Matrix ===

  a  b   <-- classified as
  6 41 |  a = POSITIVE
  8 73 |  b = NEGATIVE
```

K-STAR

```
=== Summary ===

Correctly Classified Instances          78               60.9375 %
Incorrectly Classified Instances        50               39.0625 %
Kappa statistic                          0.0294
Mean absolute error                      0.4256
Root mean squared error                  0.4861
Relative absolute error                 95.5967 %
Root relative squared error             99.7025 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,149    0,123    0,412      0,149    0,219      0,036    0,592     0,430     POSITIVE
                 0,877    0,851    0,640      0,877    0,740      0,036    0,592     0,721     NEGATIVE
Weighted Avg.    0,609    0,584    0,556      0,609    0,548      0,036    0,592     0,614

=== Confusion Matrix ===

  a  b   <-- classified as
  7 40 |  a = POSITIVE
 10 71 |  b = NEGATIVE
```

Multilayer Perceptron

Teste 2 - Parâmetros:

a) COM MERGE NOT ⇔ NOTEBOOK
b) COM HASHTAGS
c) K = **5**

```
=== Summary ===

Correctly Classified Instances          74               57.8125 %
Incorrectly Classified Instances        54               42.1875 %
Kappa statistic                         -0.0482
Mean absolute error                      0.4353
Root mean squared error                  0.5001
Relative absolute error                 97.7631 %
Root relative squared error            102.5612 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,106    0,148    0,294      0,106    0,156      -0,059   0,560     0,398     POSITIVE
                 0,852    0,894    0,622      0,852    0,719      -0,059   0,560     0,699     NEGATIVE
Weighted Avg.    0,578    0,620    0,501      0,578    0,512      -0,059   0,560     0,588

=== Confusion Matrix ===

  a  b   <-- classified as
  5 42 |  a = POSITIVE
 12 69 |  b = NEGATIVE
```

KNN

```
=== Summary ===

Correctly Classified Instances          77               60.1563 %
Incorrectly Classified Instances        51               39.8438 %
Kappa statistic                         -0.0382
Mean absolute error                      0.4394
Root mean squared error                  0.4836
Relative absolute error                 98.6916 %
Root relative squared error             99.1712 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,043    0,074    0,250      0,043    0,073      -0,063   0,570     0,410     POSITIVE
                 0,926    0,957    0,625      0,926    0,746      -0,063   0,570     0,722     NEGATIVE
Weighted Avg.    0,602    0,633    0,487      0,602    0,499      -0,063   0,570     0,607

=== Confusion Matrix ===

  a  b   <-- classified as
  2 45 |  a = POSITIVE
  6 75 |  b = NEGATIVE
```

K-STAR

```
=== Summary ===

Correctly Classified Instances          80               62.5    %
Incorrectly Classified Instances        48               37.5    %
Kappa statistic                          0.1139
Mean absolute error                      0.4297
Root mean squared error                  0.4882
Relative absolute error                 96.5042 %
Root relative squared error            100.1293 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,277    0,173    0,481      0,277    0,351      0,123    0,581     0,425     POSITIVE
                 0,827    0,723    0,663      0,827    0,736      0,123    0,581     0,719     NEGATIVE
Weighted Avg.    0,625    0,521    0,597      0,625    0,595      0,123    0,581     0,611

=== Confusion Matrix ===

  a  b   <-- classified as
 13 34 |  a = POSITIVE
 14 67 |  b = NEGATIVE
```

Multilayer Perceptron

Teste 3 - Parâmetros:

a) SEM MERGE NOT ⇔ NOTEBOOK
b) SEM HASHTAGS
c) K = **5**

```
=== Summary ===

Correctly Classified Instances          77               60.1563 %
Incorrectly Classified Instances        51               39.8438 %
Kappa statistic                          0.0349
Mean absolute error                      0.4288
Root mean squared error                  0.488
Relative absolute error                 96.3082 %
Root relative squared error            100.0869 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,191    0,160    0,409      0,191    0,261      0,040    0,583     0,428     POSITIVE
                 0,840    0,809    0,642      0,840    0,727      0,040    0,583     0,698     NEGATIVE
Weighted Avg.    0,602    0,571    0,556      0,602    0,556      0,040    0,583     0,599

=== Confusion Matrix ===

  a  b   <-- classified as
  9 38 |  a = POSITIVE
 13 68 |  b = NEGATIVE
```

KNN

```
=== Summary ===

Correctly Classified Instances          80               62.5    %
Incorrectly Classified Instances        48               37.5    %
Kappa statistic                          0.0585
Mean absolute error                      0.434
Root mean squared error                  0.4788
Relative absolute error                 97.4752 %
Root relative squared error             98.202  %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,149    0,099    0,467      0,149    0,226      0,075    0,590     0,440     POSITIVE
                 0,901    0,851    0,646      0,901    0,753      0,075    0,590     0,711     NEGATIVE
Weighted Avg.    0,625    0,575    0,580      0,625    0,559      0,075    0,590     0,611

=== Confusion Matrix ===

  a  b   <-- classified as
  7 40 |  a = POSITIVE
  8 73 |  b = NEGATIVE
```

K-STAR

```
=== Summary ===

Correctly Classified Instances          79               61.7188 %
Incorrectly Classified Instances        49               38.2813 %
Kappa statistic                          0.0537
Mean absolute error                      0.4271
Root mean squared error                  0.4854
Relative absolute error                 95.9206 %
Root relative squared error             99.5496 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,170    0,123    0,444      0,170    0,246      0,065    0,587     0,432     POSITIVE
                 0,877    0,830    0,645      0,877    0,743      0,065    0,587     0,702     NEGATIVE
Weighted Avg.    0,617    0,570    0,572      0,617    0,561      0,065    0,587     0,602

=== Confusion Matrix ===

  a  b   <-- classified as
  8 39 |  a = POSITIVE
 10 71 |  b = NEGATIVE
```

Multilayer Perceptron

Teste 4 - Parâmetros:

a) COM MERGE NOT ⇔ NOTEBOOK
b) SEM HASHTAGS
c) K = **5**

```
=== Summary ===

Correctly Classified Instances          73               57.0313 %
Incorrectly Classified Instances        55               42.9688 %
Kappa statistic                         -0.0408
Mean absolute error                      0.4311
Root mean squared error                  0.495
Relative absolute error                 96.8317 %
Root relative squared error            101.5149 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,149    0,185    0,318      0,149    0,203      -0,046   0,573     0,405     POSITIVE
                 0,815    0,851    0,623      0,815    0,706      -0,046   0,573     0,701     NEGATIVE
Weighted Avg.    0,570    0,607    0,511      0,570    0,521      -0,046   0,573     0,592

=== Confusion Matrix ===

  a  b   <-- classified as
  7 40 |  a = POSITIVE
 15 66 |  b = NEGATIVE
```

KNN

```
=== Summary ===

Correctly Classified Instances          77               60.1563 %
Incorrectly Classified Instances        51               39.8438 %
Kappa statistic                         -0.0382
Mean absolute error                      0.4367
Root mean squared error                  0.4809
Relative absolute error                 98.0872 %
Root relative squared error             98.6207 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,043    0,074    0,250      0,043    0,073      -0,063   0,587     0,424     POSITIVE
                 0,926    0,957    0,625      0,926    0,746      -0,063   0,587     0,718     NEGATIVE
Weighted Avg.    0,602    0,633    0,487      0,602    0,499      -0,063   0,587     0,610

=== Confusion Matrix ===

  a  b   <-- classified as
  2 45 |  a = POSITIVE
  6 75 |  b = NEGATIVE
```

K-STAR

```
=== Summary ===

Correctly Classified Instances          80               62.5    %
Incorrectly Classified Instances        48               37.5    %
Kappa statistic                          0.1632
Mean absolute error                      0.4217
Root mean squared error                  0.4795
Relative absolute error                 94.7121 %
Root relative squared error             98.34   %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,404    0,247    0,487      0,404    0,442      0,165    0,598     0,462     POSITIVE
                 0,753    0,596    0,685      0,753    0,718      0,165    0,598     0,714     NEGATIVE
Weighted Avg.    0,625    0,468    0,613      0,625    0,616      0,165    0,598     0,621

=== Confusion Matrix ===

  a  b   <-- classified as
 19 28 |  a = POSITIVE
 20 61 |  b = NEGATIVE
```

Multilayer Perceptron

Não houveram mudanças significativas entre as variações de *corpus*, com K = 5.

Teste 5 - Parâmetros:

d) SEM MERGE NOT ⇔ NOTEBOOK
e) COM HASHTAGS
f) K = **25**

```
=== Summary ===

Correctly Classified Instances          82               64.0625 %
Incorrectly Classified Instances        46               35.9375 %
Kappa statistic                          0.2469
Mean absolute error                      0.3835
Root mean squared error                  0.5189
Relative absolute error                 86.1322 %
Root relative squared error            106.4279 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,574    0,321    0,509      0,574    0,540      0,248    0,660     0,510     POSITIVE
                 0,679    0,426    0,733      0,679    0,705      0,248    0,660     0,739     NEGATIVE
Weighted Avg.    0,641    0,387    0,651      0,641    0,644      0,248    0,660     0,655

=== Confusion Matrix ===

  a  b   <-- classified as
 27 20 |  a = POSITIVE
 26 55 |  b = NEGATIVE
```

KNN

```
=== Summary ===

Correctly Classified Instances          82               64.0625 %
Incorrectly Classified Instances        46               35.9375 %
Kappa statistic                          0.1672
Mean absolute error                      0.3999
Root mean squared error                  0.4734
Relative absolute error                 89.8241 %
Root relative squared error             97.083  %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,340    0,185    0,516      0,340    0,410      0,175    0,664     0,562     POSITIVE
                 0,815    0,660    0,680      0,815    0,742      0,175    0,664     0,774     NEGATIVE
Weighted Avg.    0,641    0,485    0,620      0,641    0,620      0,175    0,664     0,696

=== Confusion Matrix ===

  a  b   <-- classified as
 16 31 |  a = POSITIVE
 15 66 |  b = NEGATIVE
```

K-STAR

```
=== Summary ===

Correctly Classified Instances          84               65.625  %
Incorrectly Classified Instances        44               34.375  %
Kappa statistic                          0.2669
Mean absolute error                      0.3599
Root mean squared error                  0.5315
Relative absolute error                 80.8352 %
Root relative squared error            109.01   %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,553    0,284    0,531      0,553    0,542      0,267    0,686     0,566     POSITIVE
                 0,716    0,447    0,734      0,716    0,725      0,267    0,686     0,786     NEGATIVE
Weighted Avg.    0,656    0,387    0,659      0,656    0,658      0,267    0,686     0,705

=== Confusion Matrix ===

  a  b   <-- classified as
 26 21 |  a = POSITIVE
 23 58 |  b = NEGATIVE
```

Multilayer Perceptron

Teste 6 - Parâmetros:

d) COM MERGE NOT ⇔ NOTEBOOK
e) COM HASHTAGS
f) K = **25**

```
=== Summary ===

Correctly Classified Instances          82               64.0625 %
Incorrectly Classified Instances        46               35.9375 %
Kappa statistic                          0.2403
Mean absolute error                      0.3973
Root mean squared error                  0.5387
Relative absolute error                 89.236  %
Root relative squared error            110.4818 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,553    0,309    0,510      0,553    0,531      0,241    0,630     0,506     POSITIVE
                 0,691    0,447    0,727      0,691    0,709      0,241    0,630     0,714     NEGATIVE
Weighted Avg.    0,641    0,396    0,647      0,641    0,643      0,241    0,630     0,637

=== Confusion Matrix ===

  a  b   <-- classified as
 26 21 |  a = POSITIVE
 25 56 |  b = NEGATIVE
```

KNN

```
=== Summary ===

Correctly Classified Instances          84               65.625  %
Incorrectly Classified Instances        44               34.375  %
Kappa statistic                          0.211
Mean absolute error                      0.4084
Root mean squared error                  0.484
Relative absolute error                 91.7301 %
Root relative squared error             99.2633 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,383    0,185    0,545      0,383    0,450      0,218    0,633     0,494     POSITIVE
                 0,815    0,617    0,695      0,815    0,750      0,218    0,633     0,762     NEGATIVE
Weighted Avg.    0,656    0,458    0,640      0,656    0,640      0,218    0,633     0,664

=== Confusion Matrix ===

  a  b   <-- classified as
 18 29 |  a = POSITIVE
 15 66 |  b = NEGATIVE
```

K-STAR

```
=== Summary ===

Correctly Classified Instances          77               60.1563 %
Incorrectly Classified Instances        51               39.8438 %
Kappa statistic                          0.154
Mean absolute error                      0.3946
Root mean squared error                  0.5621
Relative absolute error                 88.6286 %
Root relative squared error            115.2815 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,489    0,333    0,460      0,489    0,474      0,154    0,648     0,509     POSITIVE
                 0,667    0,511    0,692      0,667    0,679      0,154    0,648     0,781     NEGATIVE
Weighted Avg.    0,602    0,446    0,607      0,602    0,604      0,154    0,648     0,681

=== Confusion Matrix ===

  a  b   <-- classified as
 23 24 |  a = POSITIVE
 27 54 |  b = NEGATIVE
```

Multilayer Perceptron

Teste 7 - Parâmetros:

d) SEM MERGE NOT ⇔ NOTEBOOK
e) SEM HASHTAGS
f) K = **25**

```
=== Summary ===

Correctly Classified Instances          80               62.5    %
Incorrectly Classified Instances        48               37.5    %
Kappa statistic                          0.2002
Mean absolute error                      0.3979
Root mean squared error                  0.5293
Relative absolute error                 89.3712 %
Root relative squared error            108.556  %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,511    0,309    0,490      0,511    0,500      0,200    0,623     0,490     POSITIVE
                 0,691    0,489    0,709      0,691    0,700      0,200    0,623     0,709     NEGATIVE
Weighted Avg.    0,625    0,423    0,628      0,625    0,627      0,200    0,623     0,629

=== Confusion Matrix ===

  a  b   <-- classified as
 24 23 |  a = POSITIVE
 25 56 |  b = NEGATIVE
```

KNN

```
=== Summary ===

Correctly Classified Instances          81               63.2813 %
Incorrectly Classified Instances        47               36.7188 %
Kappa statistic                          0.1532
Mean absolute error                      0.4058
Root mean squared error                  0.4787
Relative absolute error                 91.145  %
Root relative squared error             98.1755 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,340    0,198    0,500      0,340    0,405      0,159    0,637     0,540     POSITIVE
                 0,802    0,660    0,677      0,802    0,734      0,159    0,637     0,756     NEGATIVE
Weighted Avg.    0,633    0,490    0,612      0,633    0,614      0,159    0,637     0,676

=== Confusion Matrix ===

  a  b   <-- classified as
 16 31 |  a = POSITIVE
 16 65 |  b = NEGATIVE
```

K-STAR

```
=== Summary ===

Correctly Classified Instances          81               63.2813 %
Incorrectly Classified Instances        47               36.7188 %
Kappa statistic                          0.2338
Mean absolute error                      0.3802
Root mean squared error                  0.55
Relative absolute error                 85.3993 %
Root relative squared error            112.7884 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,574    0,333    0,500      0,574    0,535      0,235    0,658     0,485     POSITIVE
                 0,667    0,426    0,730      0,667    0,697      0,235    0,658     0,785     NEGATIVE
Weighted Avg.    0,633    0,392    0,645      0,633    0,637      0,235    0,658     0,675

=== Confusion Matrix ===

  a  b   <-- classified as
 27 20 |  a = POSITIVE
 27 54 |  b = NEGATIVE
```

Multilayer Perceptron

Teste 8 - Parâmetros:

d) COM MERGE NOT ⇔ NOTEBOOK
e) SEM HASHTAGS
f) K = **25**

```
=== Summary ===

Correctly Classified Instances          81               63.2813 %
Incorrectly Classified Instances        47               36.7188 %
Kappa statistic                          0.2134
Mean absolute error                      0.3991
Root mean squared error                  0.5341
Relative absolute error                 89.6497 %
Root relative squared error            109.5276 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,511    0,296    0,500      0,511    0,505      0,213    0,625     0,504     POSITIVE
                 0,704    0,489    0,713      0,704    0,708      0,213    0,625     0,710     NEGATIVE
Weighted Avg.    0,633    0,418    0,634      0,633    0,634      0,213    0,625     0,634

=== Confusion Matrix ===

  a  b   <-- classified as
 24 23 |  a = POSITIVE
 24 57 |  b = NEGATIVE
```

KNN

```
=== Summary ===

Correctly Classified Instances          82               64.0625 %
Incorrectly Classified Instances        46               35.9375 %
Kappa statistic                          0.1672
Mean absolute error                      0.4096
Root mean squared error                  0.4826
Relative absolute error                 91.9933 %
Root relative squared error             98.9718 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,340    0,185    0,516      0,340    0,410      0,175    0,630     0,490     POSITIVE
                 0,815    0,660    0,680      0,815    0,742      0,175    0,630     0,757     NEGATIVE
Weighted Avg.    0,641    0,485    0,620      0,641    0,620      0,175    0,630     0,659

=== Confusion Matrix ===

  a  b   <-- classified as
 16 31 |  a = POSITIVE
 15 66 |  b = NEGATIVE
```

K-STAR

```
=== Summary ===

Correctly Classified Instances          78               60.9375 %
Incorrectly Classified Instances        50               39.0625 %
Kappa statistic                          0.1669
Mean absolute error                      0.39
Root mean squared error                  0.556
Relative absolute error                 87.5967 %
Root relative squared error            114.0188 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,489    0,321    0,469      0,489    0,479      0,167    0,665     0,534     POSITIVE
                 0,679    0,511    0,696      0,679    0,688      0,167    0,665     0,788     NEGATIVE
Weighted Avg.    0,609    0,441    0,613      0,609    0,611      0,167    0,665     0,694

=== Confusion Matrix ===

  a  b   <-- classified as
 23 24 |  a = POSITIVE
 26 55 |  b = NEGATIVE
```

Multilayer Perceptron

Não houveram mudanças significativas entre as variações de *corpus*, com K = 25.

Teste 9 - Parâmetros:

g) SEM MERGE NOT ⇔ NOTEBOOK
h) COM HASHTAGS
i) K = **50**

```
=== Summary ===

Correctly Classified Instances          84               65.625  %
Incorrectly Classified Instances        44               34.375  %
Kappa statistic                          0.2536
Mean absolute error                      0.3506
Root mean squared error                  0.5031
Relative absolute error                 78.7537 %
Root relative squared error            103.1719 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,511    0,259    0,533      0,511    0,522      0,254    0,686     0,561     POSITIVE
                 0,741    0,489    0,723      0,741    0,732      0,254    0,686     0,763     NEGATIVE
Weighted Avg.    0,656    0,405    0,653      0,656    0,655      0,254    0,686     0,689

=== Confusion Matrix ===

  a  b   <-- classified as
 24 23 |  a = POSITIVE
 21 60 |  b = NEGATIVE
```

KNN

```
=== Summary ===

Correctly Classified Instances          86               67.1875 %
Incorrectly Classified Instances        42               32.8125 %
Kappa statistic                          0.2396
Mean absolute error                      0.3783
Root mean squared error                  0.469
Relative absolute error                 84.9697 %
Root relative squared error             96.1883 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,383    0,160    0,581      0,383    0,462      0,250    0,689     0,570     POSITIVE
                 0,840    0,617    0,701      0,840    0,764      0,250    0,689     0,790     NEGATIVE
Weighted Avg.    0,672    0,449    0,657      0,672    0,653      0,250    0,689     0,709

=== Confusion Matrix ===

  a  b   <-- classified as
 18 29 |  a = POSITIVE
 13 68 |  b = NEGATIVE
```

K-STAR

```
=== Summary ===

Correctly Classified Instances          88               68.75   %
Incorrectly Classified Instances        40               31.25   %
Kappa statistic                          0.3276
Mean absolute error                      0.3155
Root mean squared error                  0.5175
Relative absolute error                 70.8601 %
Root relative squared error            106.129  %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,574    0,247    0,574      0,574    0,574      0,328    0,720     0,617     POSITIVE
                 0,753    0,426    0,753      0,753    0,753      0,328    0,720     0,796     NEGATIVE
Weighted Avg.    0,688    0,360    0,688      0,688    0,688      0,328    0,720     0,730

=== Confusion Matrix ===

  a  b   <-- classified as
 27 20 |  a = POSITIVE
 20 61 |  b = NEGATIVE
```

Multilayer Perceptron

Teste 10 - Parâmetros:

g) COM MERGE NOT ⇔ NOTEBOOK
h) COM HASHTAGS
i) K = **50**

```
=== Summary ===

Correctly Classified Instances          82               64.0625 %
Incorrectly Classified Instances        46               35.9375 %
Kappa statistic                          0.2335
Mean absolute error                      0.371
Root mean squared error                  0.5079
Relative absolute error                 83.3395 %
Root relative squared error            104.173  %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,532    0,296    0,510      0,532    0,521      0,234    0,647     0,544     POSITIVE
                 0,704    0,468    0,722      0,704    0,713      0,234    0,647     0,707     NEGATIVE
Weighted Avg.    0,641    0,405    0,644      0,641    0,642      0,234    0,647     0,647

=== Confusion Matrix ===

  a  b   <-- classified as
 25 22 |  a = POSITIVE
 24 57 |  b = NEGATIVE
```

KNN

```
=== Summary ===

Correctly Classified Instances          85               66.4063 %
Incorrectly Classified Instances        43               33.5938 %
Kappa statistic                          0.2101
Mean absolute error                      0.3925
Root mean squared error                  0.4768
Relative absolute error                 88.148  %
Root relative squared error             97.7935 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,340    0,148    0,571      0,340    0,427      0,224    0,665     0,515     POSITIVE
                 0,852    0,660    0,690      0,852    0,762      0,224    0,665     0,782     NEGATIVE
Weighted Avg.    0,664    0,472    0,646      0,664    0,639      0,224    0,665     0,684

=== Confusion Matrix ===

  a  b   <-- classified as
 16 31 |  a = POSITIVE
 12 69 |  b = NEGATIVE
```

K-STAR

```
=== Summary ===

Correctly Classified Instances          87               67.9688 %
Incorrectly Classified Instances        41               32.0313 %
Kappa statistic                          0.3258
Mean absolute error                      0.3448
Root mean squared error                  0.5185
Relative absolute error                 77.4518 %
Root relative squared error            106.3424 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,617    0,284    0,558      0,617    0,586      0,327    0,700     0,597     POSITIVE
                 0,716    0,383    0,763      0,716    0,739      0,327    0,700     0,785     NEGATIVE
Weighted Avg.    0,680    0,347    0,688      0,680    0,683      0,327    0,700     0,716

=== Confusion Matrix ===

  a  b   <-- classified as
 29 18 |  a = POSITIVE
 23 58 |  b = NEGATIVE
```

Multilayer Perceptron

Teste 11 - Parâmetros:

g) SEM MERGE NOT ⇔ NOTEBOOK
h) SEM HASHTAGS
i) K = **50**

```
=== Summary ===

Correctly Classified Instances          79               61.7188 %
Incorrectly Classified Instances        49               38.2813 %
Kappa statistic                          0.1497
Mean absolute error                      0.3706
Root mean squared error                  0.5219
Relative absolute error                 83.2386 %
Root relative squared error            107.0261 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,404    0,259    0,475      0,404    0,437      0,151    0,644     0,541     POSITIVE
                 0,741    0,596    0,682      0,741    0,710      0,151    0,644     0,732     NEGATIVE
Weighted Avg.    0,617    0,472    0,606      0,617    0,610      0,151    0,644     0,662

=== Confusion Matrix ===

  a  b   <-- classified as
 19 28 |  a = POSITIVE
 21 60 |  b = NEGATIVE
```

KNN

```
=== Summary ===

Correctly Classified Instances          85               66.4063 %
Incorrectly Classified Instances        43               33.5938 %
Kappa statistic                          0.2101
Mean absolute error                      0.3894
Root mean squared error                  0.4763
Relative absolute error                 87.4511 %
Root relative squared error             97.6775 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,340    0,148    0,571      0,340    0,427      0,224    0,663     0,547     POSITIVE
                 0,852    0,660    0,690      0,852    0,762      0,224    0,663     0,776     NEGATIVE
Weighted Avg.    0,664    0,472    0,646      0,664    0,639      0,224    0,663     0,692

=== Confusion Matrix ===

  a  b   <-- classified as
 16 31 |  a = POSITIVE
 12 69 |  b = NEGATIVE
```

K-STAR

```
=== Summary ===

Correctly Classified Instances          85               66.4063 %
Incorrectly Classified Instances        43               33.5938 %
Kappa statistic                          0.2803
Mean absolute error                      0.3543
Root mean squared error                  0.5478
Relative absolute error                 79.5775 %
Root relative squared error            112.3465 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,553    0,272    0,542      0,553    0,547      0,280    0,679     0,560     POSITIVE
                 0,728    0,447    0,738      0,728    0,733      0,280    0,679     0,763     NEGATIVE
Weighted Avg.    0,664    0,382    0,666      0,664    0,665      0,280    0,679     0,688

=== Confusion Matrix ===

  a  b   <-- classified as
 26 21 |  a = POSITIVE
 22 59 |  b = NEGATIVE
```

Multilayer Perceptron

Teste 12 - Parâmetros:

g) COM MERGE NOT ⇔ NOTEBOOK
h) SEM HASHTAGS
i) K = **50**

```
=== Summary ===

Correctly Classified Instances          79               61.7188 %
Incorrectly Classified Instances        49               38.2813 %
Kappa statistic                          0.1651
Mean absolute error                      0.3773
Root mean squared error                  0.515
Relative absolute error                 84.7449 %
Root relative squared error            105.6268 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,447    0,284    0,477      0,447    0,462      0,165    0,622     0,545     POSITIVE
                 0,716    0,553    0,690      0,716    0,703      0,165    0,622     0,698     NEGATIVE
Weighted Avg.    0,617    0,454    0,612      0,617    0,614      0,165    0,622     0,642

=== Confusion Matrix ===

  a  b   <-- classified as
 21 26 |  a = POSITIVE
 23 58 |  b = NEGATIVE
```

KNN

```
=== Summary ===

Correctly Classified Instances          85               66.4063 %
Incorrectly Classified Instances        43               33.5938 %
Kappa statistic                          0.2023
Mean absolute error                      0.3972
Root mean squared error                  0.4792
Relative absolute error                 89.2199 %
Root relative squared error             98.2755 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,319    0,136    0,577      0,319    0,411      0,220    0,657     0,508     POSITIVE
                 0,864    0,681    0,686      0,864    0,765      0,220    0,657     0,768     NEGATIVE
Weighted Avg.    0,664    0,481    0,646      0,664    0,635      0,220    0,657     0,673

=== Confusion Matrix ===

  a  b   <-- classified as
 15 32 |  a = POSITIVE
 11 70 |  b = NEGATIVE
```

K-STAR

```
=== Summary ===

Correctly Classified Instances          76               59.375  %
Incorrectly Classified Instances        52               40.625  %
Kappa statistic                          0.1486
Mean absolute error                      0.3901
Root mean squared error                  0.5687
Relative absolute error                 87.6129 %
Root relative squared error            116.6329 %
Total Number of Instances              128

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,511    0,358    0,453      0,511    0,480      0,149    0,665     0,542     POSITIVE
                 0,642    0,489    0,693      0,642    0,667      0,149    0,665     0,766     NEGATIVE
Weighted Avg.    0,594    0,441    0,605      0,594    0,598      0,149    0,665     0,684

=== Confusion Matrix ===

  a  b   <-- classified as
 24 23 |  a = POSITIVE
 29 52 |  b = NEGATIVE
```

Multilayer Perceptron

Não houveram mudanças significativas entre as variações de *corpus*, com K = 50, entretanto, ao compararmos os resultados anteriores, percebemos que há uma melhora significativa na assertividade dos algoritmos com o aumento de K.

2. Anexos

“not” vira “**notebook”**

```
TOP 50 POSITIVE Words:
    dell notebook      not      ter    compr       ir     nao      ser      nov        q     bom   melhor    quer    muit      ja    lind      vir      dar      faz       so  problem      ach      tao      ver     marc
     ano       am    agor      aqu       hp     tecl     est      cas   precis     cois    gost     nunc    cheg   atend   obrig    volt      sit      fal    troc      top      sao    parab  samsung     mesm      car
     172      136       66       43       41       37       26       25       25       23       22       18       16       16       15       13       12       12       11       11       10       10       10       10        9
       9        9        9        9        8        8        8        8        8        8        7        7        7        7        7        6        6        6        6        6        6        6        6        6        6
####################################
TOP 50 NEGATIVE Words:
    dell notebook      nao      not    compr      ter       ir  problem       so      nov      faz       ja      ser      dar      aqu   brasil    quer   comput      ver    nunc      ach        q  funcion   reclam     troc
  suport      olh      tel      car      ano      vir      vez      fic   window      lig      sab     bost      dia      voc     tecn    agor   resolv    mand     aind  carreg       qu       vc   garant      pag     vend
     374      275      148      116      107       90       80       50       40       39       39       38       36       35       34       33       32       31       28       27       25       25       25       24       22
      22       21       21       21       20       19       19       18       18       18       18       17       17       16       16       16       16       16       16       15       15       15       15       15       14
```

```
TOP 50 POSITIVE Words:
notebook     dell      ser    compr      ter       ir     est     lind      bom     dar  problem     quer       hp     marc      sit      vir  retweet     curt    atend     volt     gost     muit        q    assim      faz
  window     posi      nov     cheg    obrig  inspiron    nunc     tecl     show    troc      ano     pass        n     link     aind     driv      uso       am     ediç   escolh      fin     agor      mei   melhor  respond
      50       37       22       12       12       10        7        6        6        5        5        4        4        4        4        4        4        4        4        3        3        3        3        3        3
       3        3        3        3        3        3        2        2        2        2        2        2        2        2        2        2        2        2        2        2        2        2        2        2        2
####################################
TOP 50 NEGATIVE Words:
notebook     dell      ser    compr      ter       ir     dar  problem     nunc      nov     quer   window      vir     bost      est      ver        q    atual      ach       vc      lig   estrag   reclam     ruim     vend
     aqu       qu      tel  funcion      ano     driv     odi     memór   conect   defeit        i       gb     troc      faz    seman      fic      car      hor  program    arrum   resolv      err   carreg      vez     mand
      87       74       26       25       18       16       12       10        9        8        8        8        6        5        5        5        5        5        5        5        5        4        4        4        4
       4        4        4        4        4        4        3        3        3        3        3        3        3        3        3        3        3        3        3        3        3        3        3        3        3
```

“e” com acento (**é**) transforma-se em “**ser”**