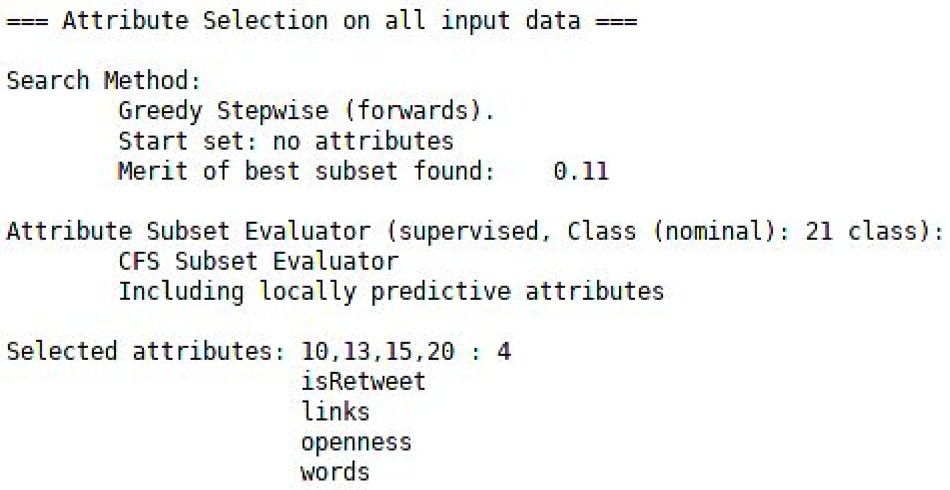
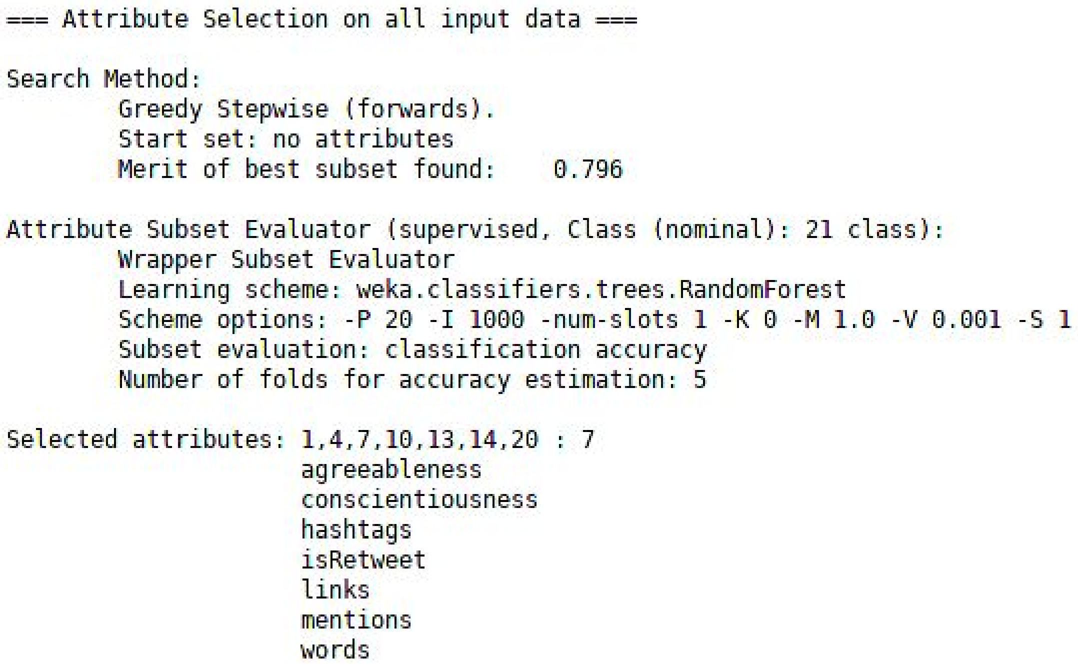
Carlos Augusto Amador Manilla
A01329447

Data Selection

Many strategies were used, the results are the following:

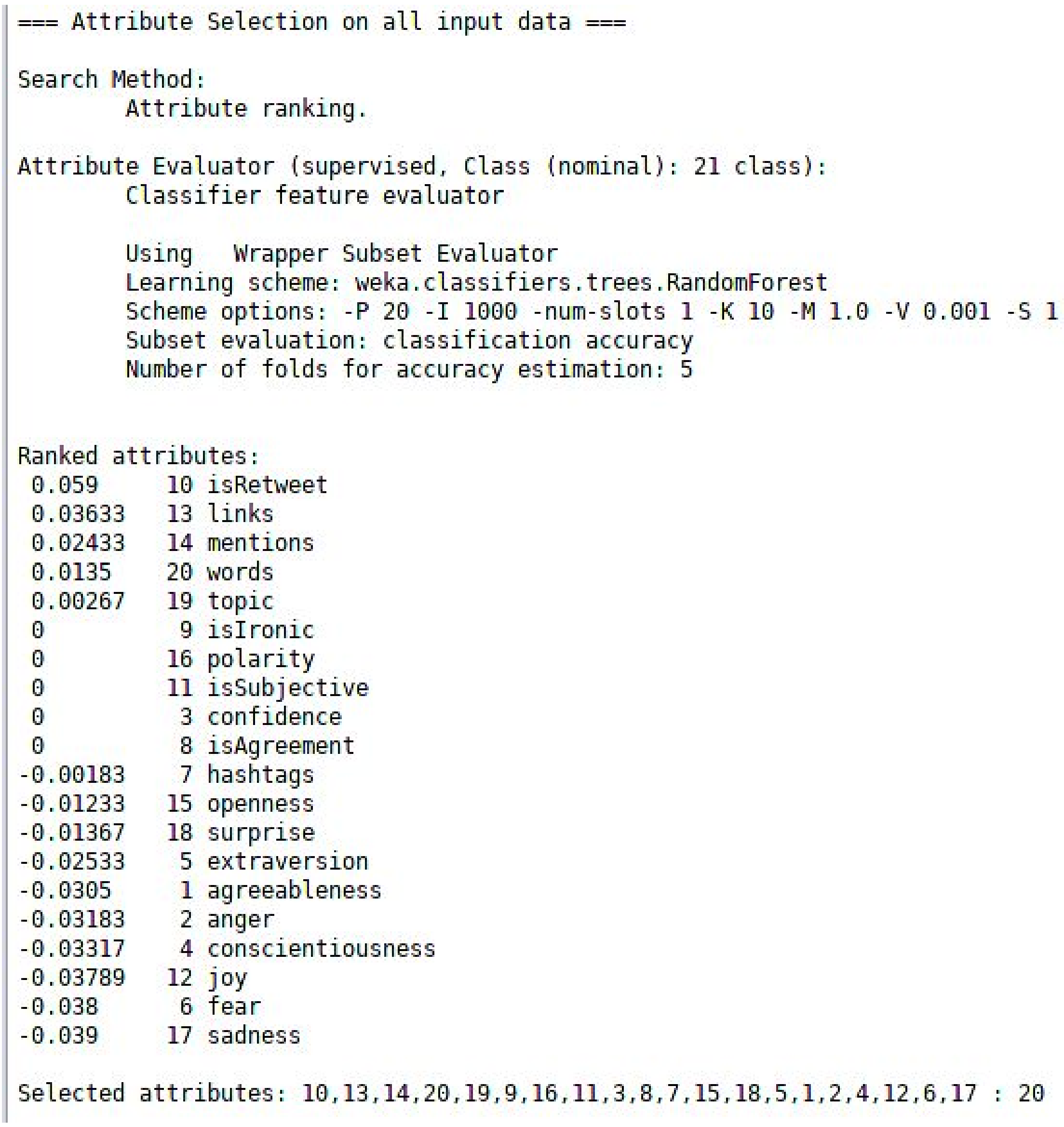


```
=== Attribute Selection on all input data ===

Search Method:
	Greedy Stepwise (forwards).
	Start set: no attributes
	Merit of best subset found:    0.11

Attribute Subset Evaluator (supervised, Class (nominal): 21 class):
	CFS Subset Evaluator
	Including locally predictive attributes

Selected attributes: 10,13,15,20 : 4
                     isRetweet
                     links
                     openness
                     words
```



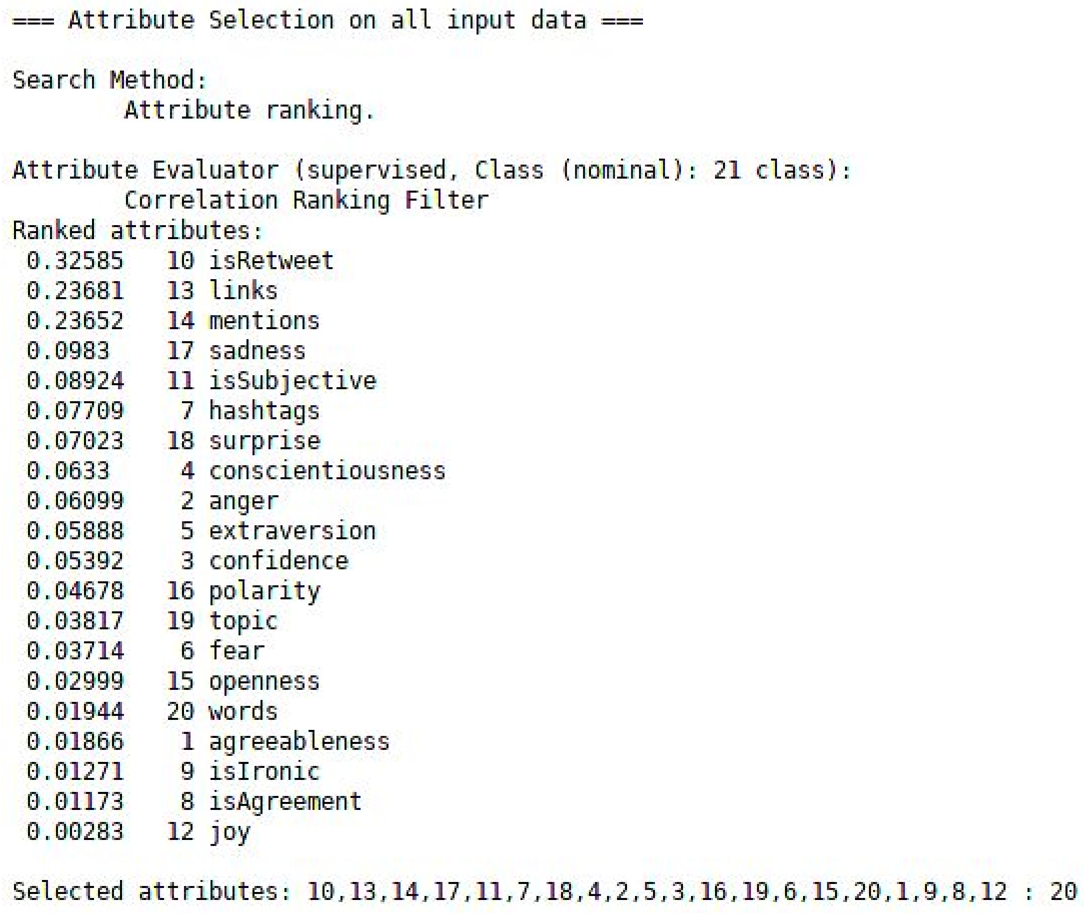
```
=== Attribute Selection on all input data ===

Search Method:
	Greedy Stepwise (forwards).
	Start set: no attributes
	Merit of best subset found:    0.796

Attribute Subset Evaluator (supervised, Class (nominal): 21 class):
	Wrapper Subset Evaluator
	Learning scheme: weka.classifiers.trees.RandomForest
	Scheme options: -P 20 -I 1000 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
	Subset evaluation: classification accuracy
	Number of folds for accuracy estimation: 5

Selected attributes: 1,4,7,10,13,14,20 : 7
                     agreeableness
                     conscientiousness
                     hashtags
                     isRetweet
                     links
                     mentions
                     words
```

Carlos Augusto Amador Manilla
A01329447



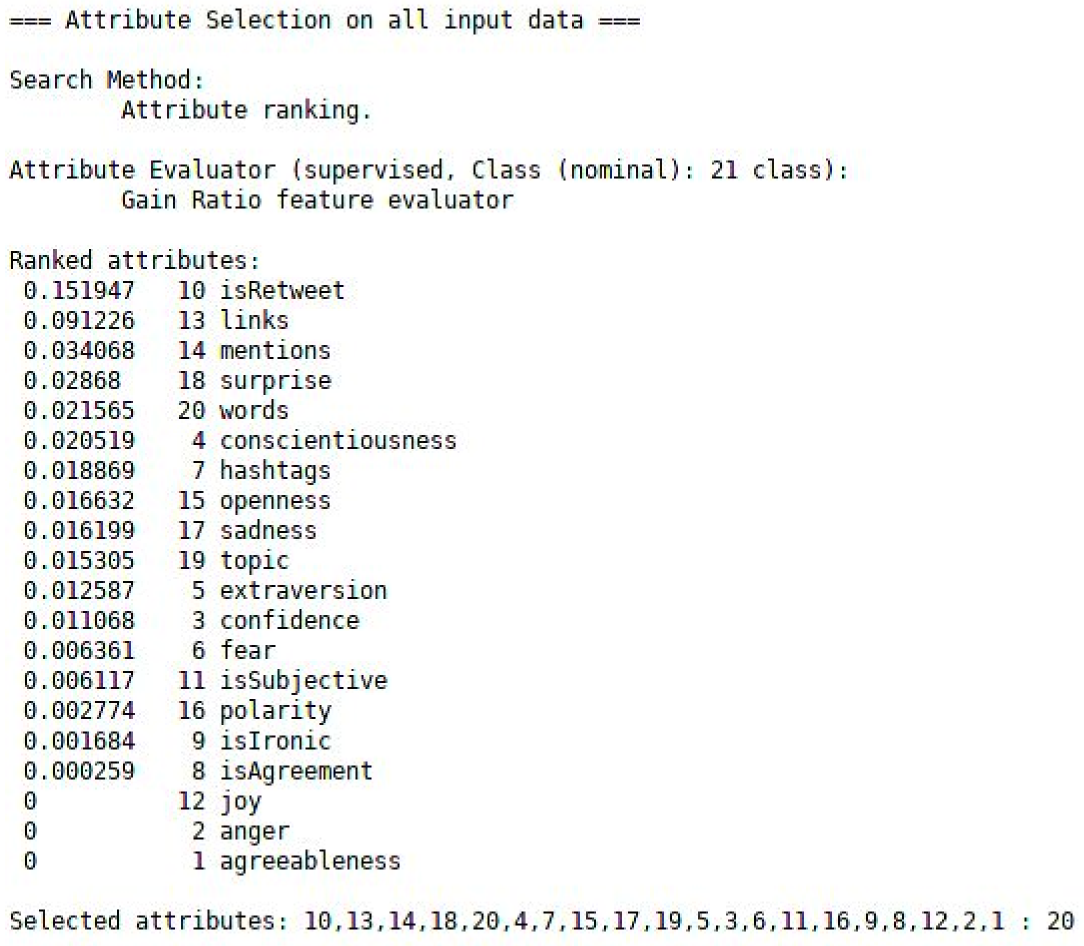
```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 21 class):
        Classifier feature evaluator

        Using   Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.trees.RandomForest
        Scheme options: -P 20 -I 1000 -num-slots 1 -K 10 -M 1.0 -V 0.001 -S 1
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5


Ranked attributes:
 0.059     10 isRetweet
 0.03633   13 links
 0.02433   14 mentions
 0.0135    20 words
 0.00267   19 topic
 0          9 isIronic
 0         16 polarity
 0         11 isSubjective
 0          3 confidence
 0          8 isAgreement
-0.00183    7 hashtags
-0.01233   15 openness
-0.01367   18 surprise
-0.02533    5 extraversion
-0.0305     1 agreeableness
-0.03183    2 anger
-0.03317    4 conscientiousness
-0.03789   12 joy
-0.038      6 fear
-0.039     17 sadness

Selected attributes: 10,13,14,20,19,9,16,11,3,8,7,15,18,5,1,2,4,12,6,17 : 20
```

Carlos Augusto Amador Manilla
A01329447



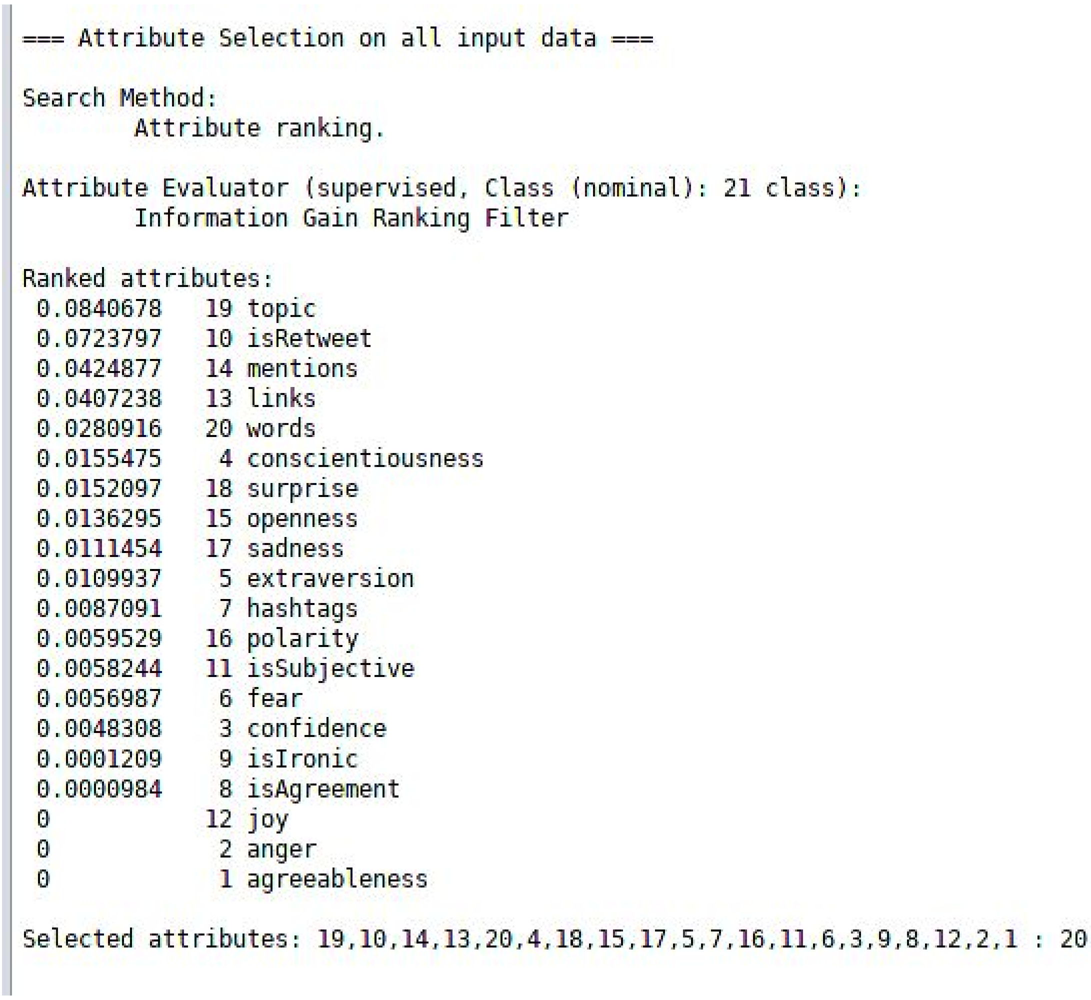
```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 21 class):
        Correlation Ranking Filter
Ranked attributes:
 0.32585   10 isRetweet
 0.23681   13 links
 0.23652   14 mentions
 0.0983    17 sadness
 0.08924   11 isSubjective
 0.07709    7 hashtags
 0.07023   18 surprise
 0.0633     4 conscientiousness
 0.06099    2 anger
 0.05888    5 extraversion
 0.05392    3 confidence
 0.04678   16 polarity
 0.03817   19 topic
 0.03714    6 fear
 0.02999   15 openness
 0.01944   20 words
 0.01866    1 agreeableness
 0.01271    9 isIronic
 0.01173    8 isAgreement
 0.00283   12 joy

Selected attributes: 10,13,14,17,11,7,18,4,2,5,3,16,19,6,15,20,1,9,8,12 : 20
```

Carlos Augusto Amador Manilla
A01329447



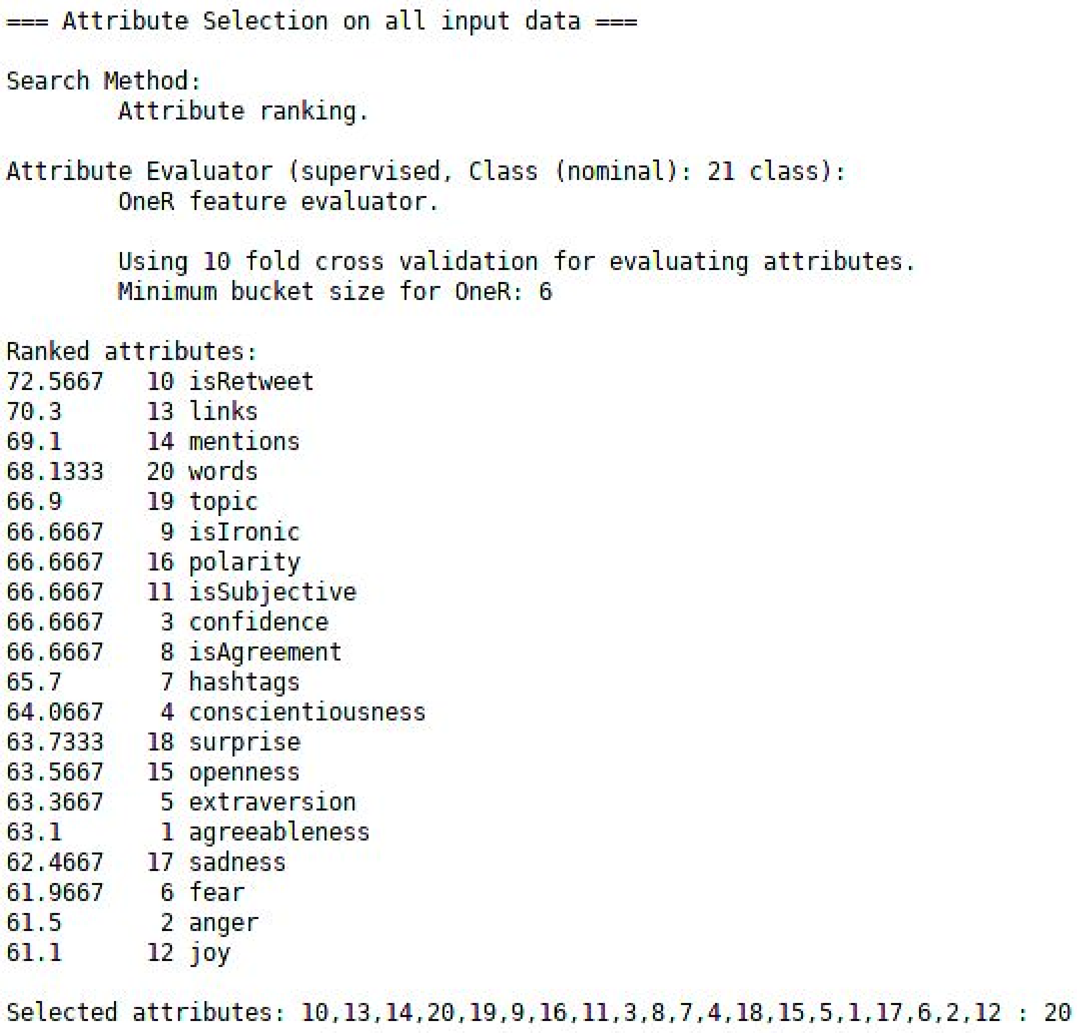
```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 21 class):
        Gain Ratio feature evaluator

Ranked attributes:
 0.151947   10 isRetweet
 0.091226   13 links
 0.034068   14 mentions
 0.02868    18 surprise
 0.021565   20 words
 0.020519    4 conscientiousness
 0.018869    7 hashtags
 0.016632   15 openness
 0.016199   17 sadness
 0.015305   19 topic
 0.012587    5 extraversion
 0.011068    3 confidence
 0.006361    6 fear
 0.006117   11 isSubjective
 0.002774   16 polarity
 0.001684    9 isIronic
 0.000259    8 isAgreement
 0          12 joy
 0           2 anger
 0           1 agreeableness

Selected attributes: 10,13,14,18,20,4,7,15,17,19,5,3,6,11,16,9,8,12,2,1 : 20
```

Carlos Augusto Amador Manilla
A01329447



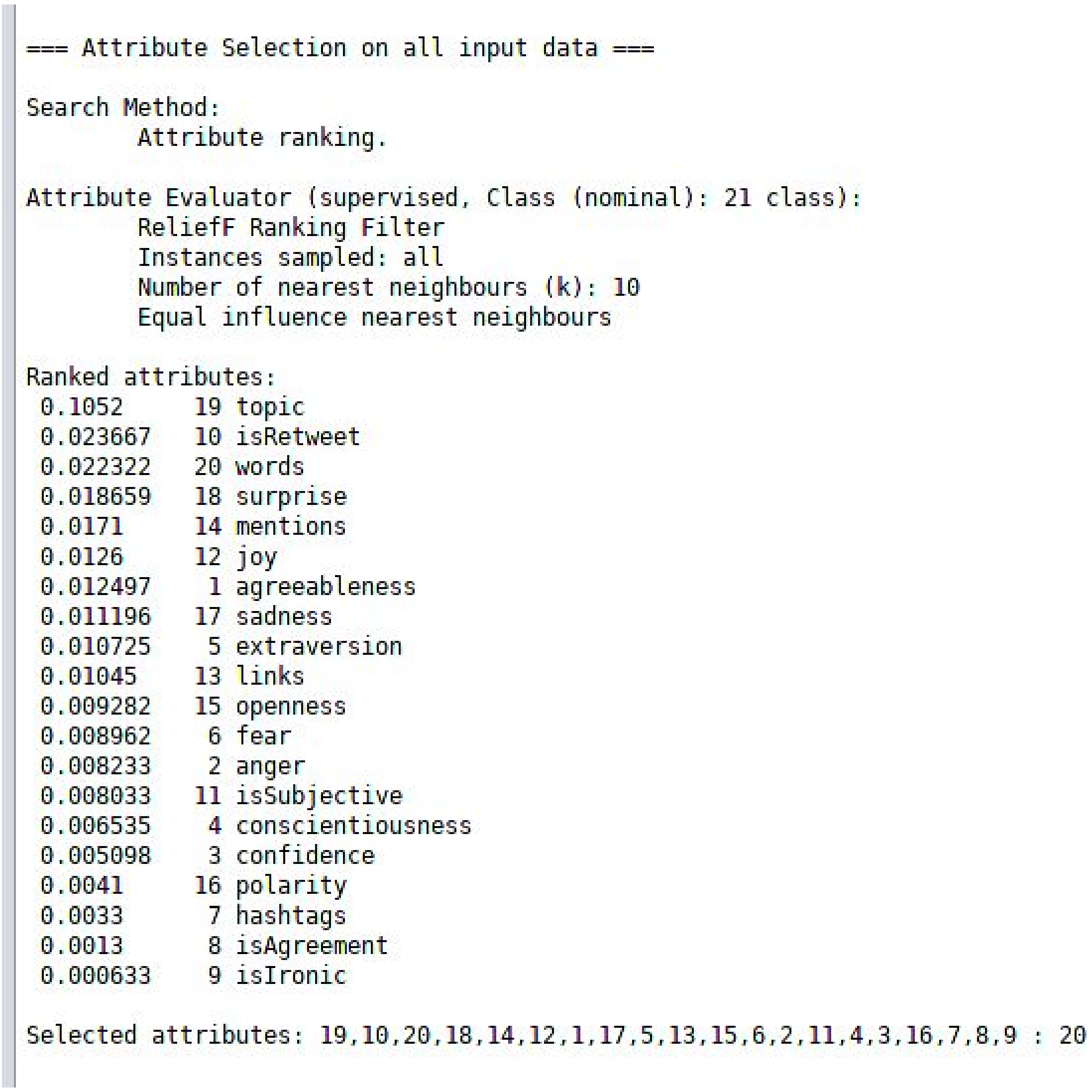
```
=== Attribute Selection on all input data ===

Search Method:
	Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 21 class):
	Information Gain Ranking Filter

Ranked attributes:
 0.0840678   19 topic
 0.0723797   10 isRetweet
 0.0424877   14 mentions
 0.0407238   13 links
 0.0280916   20 words
 0.0155475    4 conscientiousness
 0.0152097   18 surprise
 0.0136295   15 openness
 0.0111454   17 sadness
 0.0109937    5 extraversion
 0.0087091    7 hashtags
 0.0059529   16 polarity
 0.0058244   11 isSubjective
 0.0056987    6 fear
 0.0048308    3 confidence
 0.0001209    9 isIronic
 0.0000984    8 isAgreement
 0           12 joy
 0            2 anger
 0            1 agreeableness

Selected attributes: 19,10,14,13,20,4,18,15,17,5,7,16,11,6,3,9,8,12,2,1 : 20
```

Carlos Augusto Amador Manilla
A01329447



```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 21 class):
        OneR feature evaluator.

        Using 10 fold cross validation for evaluating attributes.
        Minimum bucket size for OneR: 6

Ranked attributes:
72.5667   10 isRetweet
70.3      13 links
69.1      14 mentions
68.1333   20 words
66.9      19 topic
66.6667    9 isIronic
66.6667   16 polarity
66.6667   11 isSubjective
66.6667    3 confidence
66.6667    8 isAgreement
65.7       7 hashtags
64.0667    4 conscientiousness
63.7333   18 surprise
63.5667   15 openness
63.3667    5 extraversion
63.1       1 agreeableness
62.4667   17 sadness
61.9667    6 fear
61.5       2 anger
61.1      12 joy

Selected attributes: 10,13,14,20,19,9,16,11,3,8,7,4,18,15,5,1,17,6,2,12 : 20
```

Carlos Augusto Amador Manilla
A01329447

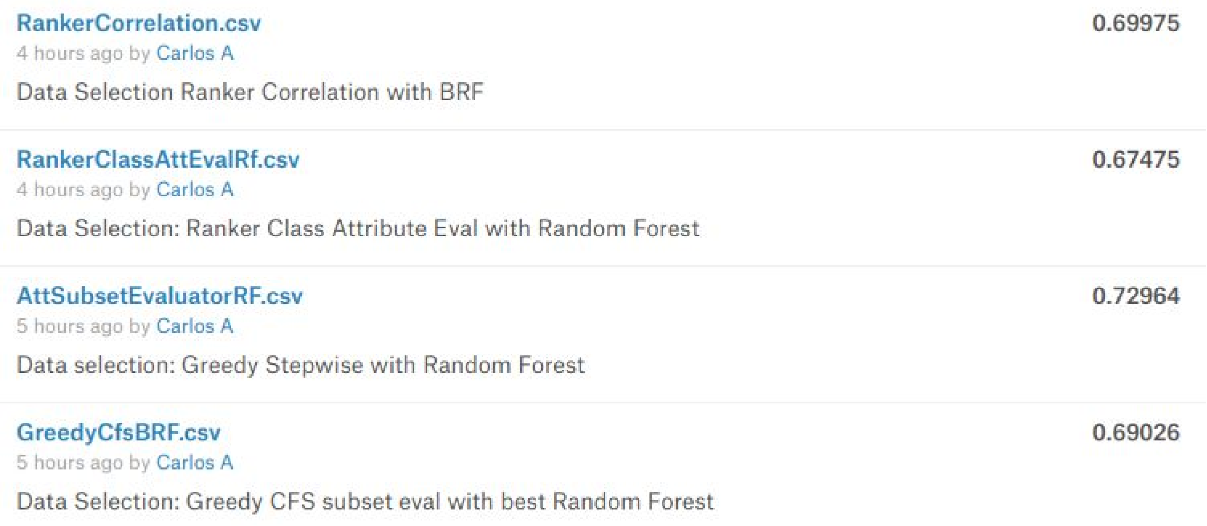


```
=== Attribute Selection on all input data ===

Search Method:
	Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 21 class):
	ReliefF Ranking Filter
	Instances sampled: all
	Number of nearest neighbours (k): 10
	Equal influence nearest neighbours

Ranked attributes:
 0.1052     19 topic
 0.023667   10 isRetweet
 0.022322   20 words
 0.018659   18 surprise
 0.0171     14 mentions
 0.0126     12 joy
 0.012497    1 agreeableness
 0.011196   17 sadness
 0.010725    5 extraversion
 0.01045    13 links
 0.009282   15 openness
 0.008962    6 fear
 0.008233    2 anger
 0.008033   11 isSubjective
 0.006535    4 conscientiousness
 0.005098    3 confidence
 0.0041     16 polarity
 0.0033      7 hashtags
 0.0013      8 isAgreement
 0.000633    9 isIronic

Selected attributes: 19,10,20,18,14,12,1,17,5,13,15,6,2,11,4,3,16,7,8,9 : 20
```

Carlos Augusto Amador Manilla
A01329447

Tests

We can observe that there are some attributes that were selected or better ranked constantly (isRetweet, words count, links count, mentions and hashtags). However there are others that introduce noise. Tests were made to check that indeed the selected attributes provided better results.



RankerCorrelation.csv 0.69975
4 hours ago by Carlos A
Data Selection Ranker Correlation with BRF

RankerClassAttEvalRf.csv 0.67475
4 hours ago by Carlos A
Data Selection: Ranker Class Attribute Eval with Random Forest

AttSubsetEvaluatorRF.csv 0.72964
5 hours ago by Carlos A
Data selection: Greedy Stepwise with Random Forest

GreedyCfsBRF.csv 0.69026
5 hours ago by Carlos A
Data Selection: Greedy CFS subset eval with best Random Forest

Conclusion

A Random Forest with only 7 attributes got better results than a Random Forest with all the attributes, this means that there are attributes which lower the accuracy of the classifier. The next step is obtain more attributes that contribute tests will be performed using data selection methods with the new attributes.