8 - Experimental Results

In this section we apply the Forward Feature selection techniques using the email corpus, and single-layer Artificial Neural Networks as classifiers.

The first step was to identify the best performance classification using ANN with Linear and Logistic Activation Function. The BPMaster program "ensayo" has the possibility to identify three types of Accuracy:

1. Training: Results (in training, validation and test sets) using the network giving the optimal training results.
2. Validation: Results (in training, validation and test sets) using the network giving the optimal validation results.
3. Test: Results (in training, validation and test sets) using the network giving the optimal test results.

For each feature we have running the classification test, using BPMaster, to evaluate the best performance for all features using Cross Validation with 5 and 10 Fold.

The algorithm run for this experimental Classification is described on the Chapter 6. The Algorithm used is on the reference Algorithm 6.3.A togheter with the Algorithm 6.3.C to obtain the following results.

The method used to make the classification experiments, was the Forward Feature Selection, using a Single-Layer ANN as classificator with double Cross-Validation using 5 Fold.

The parameters for ANN described for the experiments has been as follow:

Epoch: 400

Momentum: 0.001

Activation Function: lgt (logistic)

Weight: 0.01

Learning rate: 0.001

Bias: 0.03

Cross-Validation: 5 folds

Below you can see the results and selected features for 5 Cross-validation fold and the results for each experiment from 1 to 27 features.

Features Training by ANN with 5 CV-Fold

		Cross-Validation 5 fold	
Feature	Accuracy Train	Accuracy Validation	Accuracy Test
att_spanish	53,81%	53,90%	53,74%
att_long_words	67,14%	67,14%	67,14%
att_header_priority	67,32%	67,32%	67,32%
att_char_extend	82,27%	82,29%	82,25%
att_char_nospanish	80,55%	80,55%	80,55%
att_char_alluppercase	58,53%	58,53%	58,48%
att_spamassasin	52,41%	52,41%	52,41%
att_body_href	69,77%	69,77%	69,77%
att_crm114	62,91%	62,91%	62,91%
att_virus	55,59%	55,59%	55,59%
att_sane_spam	68,91%	68,91%	68,91%
att_sane_virus	54,36%	54,36%	54,36%
att_razor	57,73%	57,73%	57,73%
att_dcc	68,64%	68,64%	68,64%
att_vocal	50,64%	50,64%	50,64%
att_char3_number	81,65%	81,66%	81,65%
att_char_header_consonante	52,02%	52,02%	52,04%
att_header_html	74,68%	74,68%	74,68%
att_mx_sender	51,86%	51,86%	51,86%
att_body_href_img	65,73%	65,73%	65,73%
att_body_color	72,77%	72,77%	72,77%
att_body_href_extend_char	75,23%	75,23%	75,23%
att_dkim	51,91%	51,91%	51,91%
att_count_url	51,86%	51,86%	51,86%
att_spf	51,85%	51,86%	51,86%
att_pyzor	49,98%	50,27%	49,90%
att_char_from_consonante	50,04%	49,85%	50,02%

Artificial Neural Networks definition parameters

Attribute	epochs	momentum	activation function	cv-fold	weight rate	learning rate	bias	Accuracy
att_spanish	400	0.001	lgt	5	0.01	0.001	0.03	53,90%
att_long_words	400	0.001	lgt	5	0.01	0.001	0.03	67,14%
att_header_priority	400	0.001	lgt	5	0.01	0.001	0.03	67,32%
att_char_extend	400	0.001	lgt	5	0.01	0.001	0.03	82,29%
att_char_nospanish	400	0.001	lgt	5	0.01	0.001	0.03	80,55%

att_char_alluppercase	400	0.001	lgt	5	0.01	0.001	0.03	58,53%
att_spamassasin	400	0.001	lgt	5	0.01	0.001	0.03	52,41%
att_body_href	400	0.001	lgt	5	0.01	0.001	0.03	69,77%
att_crm114	400	0.001	lgt	5	0.01	0.001	0.03	62,91%
att_virus	400	0.001	lgt	5	0.01	0.001	0.03	55,59%
att_sane_spam	400	0.001	lgt	5	0.01	0.001	0.03	68,91%
att_sane_virus	400	0.001	lgt	5	0.01	0.001	0.03	54,36%
att_razor	400	0.001	lgt	5	0.01	0.001	0.03	57,73%
att_dcc	400	0.001	lgt	5	0.01	0.001	0.03	68,64%
att_vocal	400	0.001	lgt	5	0.01	0.001	0.03	50,64%
att_char3_number	400	0.001	lgt	5	0.01	0.001	0.03	81,66%
att_char_header_consonante	400	0.001	lgt	5	0.01	0.001	0.03	52,02%
att_header_html	400	0.001	lgt	5	0.01	0.001	0.03	74,68%
att_mx_sender	400	0.001	lgt	5	0.01	0.001	0.03	51,86%
att_body_href_img	400	0.001	lgt	5	0.01	0.001	0.03	65,73%
att_body_color	400	0.001	lgt	5	0.01	0.001	0.03	72,77%
att_body_href_extend_char	400	0.001	lgt	5	0.01	0.001	0.03	75,23%
att_dkim	400	0.001	lgt	5	0.01	0.001	0.03	51,91%
att_count_url	400	0.001	lgt	5	0.01	0.001	0.03	51,86%
att_spf	400	0.001	lgt	5	0.01	0.001	0.03	51,86%
att_pyzor	400	0.001	lgt	5	0.01	0.001	0.03	50,27%
att_char_from_consonante	400	0.001	lgt	5	0.01	0.001	0.03	49,85%

In the first part of the algorithm gets the value of each FS (Feature), and the accuracy itself, showing that not all FS (Features) have the same accuracy.

As the proposed objective was to obtain a 95% or more accurately, using all the features of the data set using the ANN based classifier, and using linear or logistic algorithms. The next step is to experiment with every feature to end all the features and performance rating.

Results using 1 feature – 82,25%

Attribute	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	Ensayo1	82,27%	82,29%	82,25%

Results using 2 features – 82,46%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%	Ensayo2	82,46%	82,47%	82,46%
att_char3_number	81,87%				

Results using 3 features – 82,49%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%	Ensayo3	82,49%	82,50%	82,49%
att_char3_number	81,87%				
att_char_nospanish	80,05%				

Results using 4 features – 82,35%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%	Ensayo4	82,46%	82,35%	82,35%
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				

Results using 5 features – 82,76%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%	Ensayo5	82,84%	82,79%	82,76%
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				
att_body_color	72,77%				

Results using 6 features – 82,79%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%	Ensayo6	82,99%	82,79%	82,75%
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				
att_body_color	72,77%				
att_body_href	69,75%				

Results using 7 features – 84,32%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%	Ensayo7	84,64%	84,32%	84,32%
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				
att_body_color	72,77%				
att_body_href	69,75%				
att_sane_spam	68,92%				

Results using 8 features – 84,26%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%	Ensayo8			
att_char3_number	81,87%				
att_char_nospanish	80,05%				

att_body_href_extend_char	75,23%				
att_body_color	72,77%		84,62%	84,27%	84,26%
att_body_href	69,75%				
att_sane_spam	68,92%				
att_dcc	68,64%				

Results using 9 features – 85,39%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%				
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				
att_body_color	72,77%	Ensayo9	86,11%	85,40%	85,39%
att_body_href	69,75%				
att_sane_spam	68,92%				
att_dcc	68,64%				
att_long_words	68,57%				

Results using 10 features – 86,86%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%				
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				
att_body_color	72,77%				
att_body_href	69,75%	Ensayo10	87,41%	86,86%	86,86%
att_sane_spam	68,92%				
att_dcc	68,64%				
att_long_words	68,57%				
att_header_priority	67,32%				

Results using 11 features – 88,35%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%	Ensayo11	88,51%	88,35%	88,35%
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				
att_body_color	72,77%				
att_body_href	69,75%				
att_sane_spam	68,92%				
att_dcc	68,64%				
att_long_words	68,57%				
att_header_priority	67,32%				
att_body_href_img	65,73%				

Results using 12 features – 89,38%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%	Ensayo12	89,71%	89,40%	89,38%
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				
att_body_color	72,77%				
att_body_href	69,75%				
att_sane_spam	68,92%				
att_dcc	68,64%				
att_long_words	68,57%				
att_header_priority	67,32%				
att_body_href_img	65,73%				
att_crm114	62,91%				

Results using 13 features – 91,45%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%	Ensayo13	92,12%	91,46%	91,45%
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				
att_body_color	72,77%				
att_body_href	69,75%				
att_sane_spam	68,92%				
att_dcc	68,64%				
att_long_words	68,57%				
att_header_priority	67,32%				
att_body_href_img	65,73%				
att_crm114	62,91%				
att_header_html	62,91%				

Results using 14 features – 94,83%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%	Ensayo14	95,08%	94,84%	94,83%
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				
att_body_color	72,77%				
att_body_href	69,75%				
att_sane_spam	68,92%				
att_dcc	68,64%				
att_long_words	68,57%				
att_header_priority	67,32%				
att_body_href_img	65,73%				
att_crm114	62,91%				
att_header_html	62,91%				
att_char_alluppercase	58,53%				

Results using 15 features – 94,75%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%	Ensayo15	95,01%	94,76%	94,75%
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				
att_body_color	72,77%				
att_body_href	69,75%				
att_sane_spam	68,92%				
att_dcc	68,64%				
att_long_words	68,57%				
att_header_priority	67,32%				
att_body_href_img	65,73%				
att_crm114	62,91%				
att_header_html	62,91%				
att_char_alluppercase	58,53%				
att_razor	57,73%				

Results using 16 features – 95,96%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%	Ensayo16	96,34%	95,96%	95,96%
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				
att_body_color	72,77%				
att_body_href	69,75%				
att_sane_spam	68,92%				
att_dcc	68,64%				
att_long_words	68,57%				
att_header_priority	67,32%				
att_body_href_img	65,73%				

att_crm114	62,91%	
att_header_html	62,91%	
att_char_alluppercase	58,53%	
att_razor	57,73%	
att_virus	55,59%	

Results using 17 features – 95,93%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%	Ensayo17	96,30%	95,94%	95,93%
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				
att_body_color	72,77%				
att_body_href	69,75%				
att_sane_spam	68,92%				
att_dcc	68,64%				
att_long_words	68,57%				
att_header_priority	67,32%				
att_body_href_img	65,73%				
att_crm114	62,91%				
att_header_html	62,91%				
att_char_alluppercase	58,53%				
att_razor	57,73%				
att_virus	55,59%				
att_sane_virus	54,36%				

Results using 18 features – 96,06%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%	Ensayo18			
att_char3_number	81,87%				
att_char_nospanish	80,05%				

att_body_href_extend_char	75,23%				
att_body_color	72,77%				
att_body_href	69,75%				
att_sane_spam	68,92%				
att_dcc	68,64%				
att_long_words	68,57%		96,44%	96,06%	96,06%
att_header_priority	67,32%				
att_body_href_img	65,73%				
att_crm114	62,91%				
att_header_html	62,91%				
att_char_alluppercase	58,53%				
att_razor	57,73%				
att_virus	55,59%				
att_sane_virus	54,36%				
att_spanish	53,90%				

Results using 19 features – 96,22%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%				
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				
att_body_color	72,77%				
att_body_href	69,75%				
att_sane_spam	68,92%	Ensayo19			
att_dcc	68,64%				
att_long_words	68,57%				
att_header_priority	67,32%				
att_body_href_img	65,73%		96,70%	96,23%	96,22%
att_crm114	62,91%				
att_header_html	62,91%				

att_char_alluppercase	58,53%	
att_razor	57,73%	
att_virus	55,59%	
att_sane_virus	54,36%	
att_spanish	53,90%	
att_spamassasin	52,41%	

Results using 20 features – 96,21%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%	Ensayo20	96,72%	96,21%	96,21%
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				
att_body_color	72,77%				
att_body_href	69,75%				
att_sane_spam	68,92%				
att_dcc	68,64%				
att_long_words	68,57%				
att_header_priority	67,32%				
att_body_href_img	65,73%				
att_crm114	62,91%				
att_header_html	62,91%				
att_char_alluppercase	58,53%				
att_razor	57,73%				
att_virus	55,59%				
att_sane_virus	54,36%				
att_spanish	53,90%				
att_spamassasin	52,41%				
att_char_header_consonante	52,02%				

Results using 21 features – 96,18%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%	Ensayo21	96,67%	96,21%	96,18%
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				
att_body_color	72,77%				
att_body_href	69,75%				
att_sane_spam	68,92%				
att_dcc	68,64%				
att_long_words	68,57%				
att_header_priority	67,32%				
att_body_href_img	65,73%				
att_crm114	62,91%				
att_header_html	62,91%				
att_char_alluppercase	58,53%				
att_razor	57,73%				
att_virus	55,59%				
att_sane_virus	54,36%				
att_spanish	53,90%				
att_spamassasin	52,41%				
att_char_header_consonante	52,02%				
att_dkim	51,91%				

Results using 22 features – 96,19%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%	Ensayo22			
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				

att_body_color	72,77%				
att_body_href	69,75%				
att_sane_spam	68,92%				
att_dcc	68,64%				
att_long_words	68,57%				
att_header_priority	67,32%				
att_body_href_img	65,73%				
att_crm114	62,91%				
att_header_html	62,91%				
att_char_alluppercase	58,53%				
att_razor	57,73%		96,74%	96,19%	96,19%
att_virus	55,59%				
att_sane_virus	54,36%				
att_spanish	53,90%				
att_spamassasin	52,41%				
att_char_header_consonante	52,02%				
att_dkim	51,91%				
att_count_url	51,86%				

Results using 23 features – 96,21%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%				
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				
att_body_color	72,77%				
att_body_href	69,75%	Ensayo23			
att_sane_spam	68,92%				
att_dcc	68,64%				
att_long_words	68,57%				
att_header_priority	67,32%				

att_body_href_img	65,73%				
att_crm114	62,91%				
att_header_html	62,91%		96,76%	96,21%	96,21%
att_char_alluppercase	58,53%				
att_razor	57,73%				
att_virus	55,59%				
att_sane_virus	54,36%				
att_spanish	53,90%				
att_spamassasin	52,41%				
att_char_header_consonante	52,02%				
att_dkim	51,91%				
att_count_url	51,86%				
att_mx_sender	51,86%				

Results using 24 features – 96,18%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%				
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				
att_body_color	72,77%				
att_body_href	69,75%				
att_sane_spam	68,92%				
att_dcc	68,64%	Ensayo24			
att_long_words	68,57%				
att_header_priority	67,32%				
att_body_href_img	65,73%				
att_crm114	62,91%		96,70%	96,20%	96,18%
att_header_html	62,91%				
att_char_alluppercase	58,53%				
att_razor	57,73%				

att_virus	55,59%	
att_sane_virus	54,36%	
att_spanish	53,90%	
att_spamassasin	52,41%	
att_char_header_consonante	52,02%	
att_dkim	51,91%	
att_count_url	51,86%	
att_mx_sender	51,86%	
att_spf	51,86%	

Results using 25 features – 96,19%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%	Ensayo25	96,76%	96,19%	96,19%
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				
att_body_color	72,77%				
att_body_href	69,75%				
att_sane_spam	68,92%				
att_dcc	68,64%				
att_long_words	68,57%				
att_header_priority	67,32%				
att_body_href_img	65,73%				
att_crm114	62,91%				
att_header_html	62,91%				
att_char_alluppercase	58,53%				
att_razor	57,73%				
att_virus	55,59%				
att_sane_virus	54,36%				
att_spanish	53,90%				
att_spamassasin	52,41%				
att_char_header_consonante	52,02%				

att_dkim	51,91%	
att_count_url	51,86%	
att_mx_sender	51,86%	
att_spf	51,86%	
att_char_extend	82,29%	

Results using 26 features – 96,15%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%				
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				
att_body_color	72,77%				
att_body_href	69,75%				
att_sane_spam	68,92%				
att_dcc	68,64%		96,73%	96,17%	96,15%
att_long_words	68,57%				
att_header_priority	67,32%				
att_body_href_img	65,73%				
att_crm114	62,91%				
att_header_html	62,91%	Ensayo26			
att_char_alluppercase	58,53%				
att_razor	57,73%				
att_virus	55,59%				
att_sane_virus	54,36%				
att_spanish	53,90%				
att_spamassasin	52,41%				
att_char_header_consonante	52,02%				
att_dkim	51,91%				
att_count_url	51,86%				
att_mx_sender	51,86%				
att_spf	51,86%				

att_vocal	50,64%	
att_char_from_consonante	49,85%	

Results using 27 features – 96,10%

Attribute	Accuracy	Ensayo	Accuracy Train	Accuracy Validation	Accuracy Test
att_char_extend	82,29%				
att_char3_number	81,87%				
att_char_nospanish	80,05%				
att_body_href_extend_char	75,23%				
att_body_color	72,77%				
att_body_href	69,75%				
att_sane_spam	68,92%				
att_dcc	68,64%				
att_long_words	68,57%				
att_header_priority	67,32%				
att_body_href_img	65,73%				
att_crm114	62,91%				
att_header_html	62,91%				
att_char_alluppercase	58,53%	Ensayo27			
att_razor	57,73%		96,73%	96,10%	96,10%
att_virus	55,59%				
att_sane_virus	54,36%				
att_spanish	53,90%				
att_spamassasin	52,41%				
att_char_header_consonante	52,02%				
att_dkim	51,91%				
att_count_url	51,86%				
att_mx_sender	51,86%				
att_spf	51,86%				
att_vocal	50,64%				
att_pyzor	49,90%				
att_char_from_consonante	49,85%				