Predicting Information Credibility in Time-Sensitive Social Media

Carlos Castillo, Marcelo Mendoza, Barbara Poblete

Supplementary Material

(To be available on-line)

Predicting Information Credibility in Time-Sensitive Social Media
Supplementary Material

1. Handling of the UNSURE class for newsworthiness

UNSURE labels represent cases where Mechanical Turk evaluators do not agree on the label of the case. Intuitively, UNSURE cases were ambiguous or the evidence presented to evaluators was not enough to take a decision about its scope.

We explore several alternatives to address this problem. We perform a comparison among classifiers trained from: a) Full collection of cases (NEWS/CHAT/UNSURE), b) Merged collection of cases (NEWS versus THE REST), and c) Pruned collection of cases (NEWS/CHAT). We study this problem for a fixed classifier. A preliminary evaluation conducted over the full collection of cases indicates to us that the use of a J48 decision tree achieves very competitive results with other classifiers such as naive Bayes or support vector machines. We decide to use J48 in this first evaluation, postponing the impact of model selection for a posterior evaluation.

As an evaluation strategy we consider cross validation to avoid that a particular unrepresentative training set biased the learning phase. Thus, we perform several training/testing processes, and performance measures represent averages across folds. As we have few data instances, we use leave-one-out cross-validation. In this strategy, each instance is turn left out, and the learning method is trained on all the remaining instances. This strategy offers two main advantages. First, the greatest possible amount of data is used for training in each case, which presumably increases the chance that the classifier achieves good prediction results. Second, we don't need to perform a sampling process as in n-fold cross validation methods. Overall experimental results are shown in Table 1.

As Table 1 shows, the inclusion of UNSURE labels significantly decreases the performance of the classifier. When UNSURE cases are merged with CHAT cases into one class, the accuracy achieves its highest value for this experiment. However, Table 1 shows that the lowest error measures values are achieved when UNSURE cases are dropped from the data collection. We can observe also that the biggest Kappa statistic value is achieved when UNSURE cases are eliminated. This indicates to us that the predictability of the problem achieves significant improvements when UNSURE cases are dropped. On the other hand, the inclusion of UNSURE cases in the learning phase reaches only a Kappa statistic equals to 0.1066, which indicates that improvements over a random predictor are marginal. Detailed results disaggregated per class are illustrated in Table 2.

Table 1: Results summary for the labels set study over newsworthy detection.

	NEWS/CHAT/UNSURE	NEWS/REST	NEWS/CHAT
Correctly Class. Instances	40.73%	67.36%	67.21%
Incorrectly Class. Instances	59.26%	32.63%	32.79%
Kappa statistic	0.1066	0.2134	0.338
Mean absolute error	0.3944	0.3495	0.323
Root mean squared error	0.6011	0.5579	0.542
Relative absolute error	88.81%	83.72%	64.73%
Root relative squared error	127.39%	122.01%	108.38%
Total number of instances	383	383	247

Table 2: Detailed results for the labels set study over newsworthy detection.

	NEWS/CHAT/UNSURE				
	FP Rate	Precision	Recall	F-Measure	ROC Area
NEWS	0.207	0.477	0.451	0.464	0.645
CHAT	0.277	0.444	0.41	0.426	0.552
UNSURE	0.413	0.329	0.368	0.347	0.526
Weighted Avg.	0.305	0.413	0.407	0.409	0.57
	NEWS/REST				
	FP Rate	Precision	Recall	F-Measure	ROC Area
NEWS	0.23	0.446	0.442	0.444	0.538
REST	0.558	0.768	0.77	0.769	0.538
Weighted Avg.	0.461	0.673	0.674	0.673	0.538
	NEWS/CHAT				
	FP Rate	Precision	Recall	F-Measure	ROC Area
NEWS	0.291	0.645	0.628	0.637	0.737
CHAT	0.372	0.693	0.709	0.701	0.737
Weighted Avg.	0.335	0.671	0.672	0.672	0.737

Table 2 shows that in the three classes problem, the separation of the UNSURE class is very difficult. In fact, for the UNSURE class the false positive rate achieves its highest value. Notice also that the precision achieves its lowest value. In the NEWS/REST problem, the inclusion of UNSURE cases into the REST class decreases the performance of the classifier, achieving the highest false positive rate of the experiment. A good balance is reached when UNSURE cases are dropped and the problem is reduced to NEWS/CHAT separation, where the F-measure for the NEWS class reaches its highest value. Then we reduce the newsworthy detection problem to a binary classification problem between NEWS/CHAT classes.

2. Choice of a learning scheme for newsworthiness

We tried a number of learning schemes to evaluate newsworthy detection. Learning schemes were selected from different families of machine learning approaches. We decide to explore how Bayesian methods perform in this problem, proving a naive Bayes algorithm and a Bayes network algorithm. We explore also a regression-based approach which use logistic functions. Previous methods offer to us output scores, so we can decide when the classifier identifies enough evidence to perform its predictions. We use also trees-based methods considering a random forest algorithm and a J48 classifier, whose results were shown previously in tables 1 and 2. Overall experimental results are shown in Table 3.

Table 3: Results summary for different learning algorithms over newsworthy detection.

	Naive Bayes	Bayes Net	Logistic	Random Forest
Correctly Class. Instances	50.2024%	80.1619%	70.4453%	77.7328%
Incorrectly Class. Instances	49.7976%	19.8381%	29.5547%	22.2672%
Kappa statistic	0.0612	0.5984	0.4034	0.5517
Mean absolute error	0.498	0.1989	0.3511	0.304
Root mean squared error	0.7057	0.4279	0.4903	0.404
Relative absolute error	99.9116%	39.9131%	70.441%	61.0029%
Root relative squared error	141.0788%	85.5381%	98.0153%	80.7636%

We can observe that the feature independence assumption of the Naive Bayes method achieves very poor performance results which are significantly outperformed by the Bayesian network. In fact, the accuracy of the Bayes Net is the highest of the experiment. Very close in performance to Bayes Net is the Random Forest classifier, whose root mean squared error and root relative squared error are the lowest in the experiment. An accuracy equals to 70% is reached by the logistic regression model, being the main advantage of this method its simplicity because the model is codified using only a simple weighted vector. The lowest mean absolute error is reached by the Bayes network. Very significant values for the Kappa statistics were achieved by these methods, except the Naive Bayes, indicating to us that the predictability of our classifiers are significantly better than a random predictor.

Table 4 shows how these methods performs in each class. The last row of each evaluation shows the weighted average between both classes.

As Table 4 shows, for the Naive Bayes method it was very difficult the detection of NEWS cases, leading to an incredible high false positive rate greater than 0.8. Higher precisions were achieved for NEWS detection using Random Forest and Bayes Net. CHAT detection results were very competitive for logistic regression, only a few percentage points under Random Forest and Bayes Net results.

Table 4: Detailed results for different learning algorithms over newsworthy detection.

			Naive Bayes		
	FP Rate	Precision	Recall	F-Measure	ROC Area
NEWS	0.828	0.476	0.894	0.622	0.581
CHAT	0.106	0.657	0.172	0.272	0.707
Weighted Avg.	0.437	0.574	0.502	0.432	0.649
			Bayes Net		
	FP Rate	Precision	Recall	F-Measure	ROC Area
NEWS	0.157	0.802	0.752	0.776	0.861
CHAT	0.248	0.801	0.843	0.822	0.861
Weighted Avg.	0.206	0.802	0.802	0.801	0.861
			Logistic		
	FP Rate	Precision	Recall	F-Measure	ROC Area
NEWS	0.261	0.682	0.664	0.673	0.7
CHAT	0.336	0.723	0.739	0.731	0.7
Weighted Avg.	0.302	0.704	0.704	0.704	0.7
			Random Forest		
	FP Rate	Precision	Recall	F-Measure	ROC Area
NEWS	0.209	0.754	0.761	0.758	0.839
CHAT	0.239	0.797	0.791	0.794	0.839
Weighted Avg.	0.225	0.778	0.777	0.777	0.839

ROC analysis is a very valuable analysis tool for this problem. We are interested in the detection of NEWS labels, discarding cases labeled as CHAT. Thus, the compromise between hit rate and false alarm rate characterize a trade-off which illustrate to us the NEWS detection ability of the studied methods. ROC curves illustrate this trade-off. In particular, in Table 4 we show the area under each ROC curve, which corresponds to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. Thus, higher values indicate better performance. Best values for ROC areas were achieved by Bayes Net and Random Forest, indicating to us that these methods are the best evaluated in this experiment.

3. Effects of dimensionality reduction on the newsworthy classifier

Several features show good properties regarding class separability. For example, the feature "fraction of authors with a description" indicates that more users with a description in their profiles tend to spread chat tweets. On the other hand, the feature "average length of a tweet" shows that long tweets are more related to newsworthy topics. Several interesting relations arise from these comparisons. Some of them are very intuitive, for example, tweets related to newsworthy events tend

to share URLs over the top-100 most popular domains and chat tweets tend to contain more frown emoticons. Other relations show less obvious properties such as tweets related to chat trends are related to more distinct hashtags indicating to us that newsworthy events are in general related to few hashtags. Another interesting relation indicates that newsworthy tweets tend to refer more URLs, illustrating that newsworthy events are in general reported by several news media sources.

We also study the impact of dimensionality reduction on classification performance. Results are included in Appendix 3.

To do this we reduce the feature space to the 8 best features previously discussed and we train newsworthy detection models. Then we evaluate the performance by following a leave-one-out strategy. We study a number of learning algorithms techniques, among them naive Bayes, Bayes networks, logistic regression, random forest and J48 decision trees. The poorest performance result was reached by naive Bayes, with an accuracy equals to 60%. Among the remaining four machine learning techniques, logistic regression and Bayes networks achieves very similar results. Table 5 shows these measures. We omit naive Bayes measures from the table.

Table 5: Results summary for newsworthy detection using feature selection with different learning algorithms.

	Bayes Net	Logistic	Random forest	J48
Correctly Class. Instances	75.3036%	75.3036%	74.0891%	68.8259%
Incorrectly Class. Instances	24.6964%	24.6964%	25.9109%	31.1741%
Kappa statistic	0.4951	0.5014	0.4802	0.3636
Mean absolute error	0.263	0.3555	0.3142	0.329
Root mean squared error	0.412	0.4289	0.4339	0.4598
Relative absolute error	52.7592%	71.3334%	63.0336%	66.0024%
Root relative squared error	82.3745%	85.7484%	86.7432%	91.9284%

Table 5 shows that Bayes networks and logistic regression results achieves the same number of correctly classified instances but there are some differences in error measures, being in general error measures results more strong for the Bayesian-based method. Logistic regression achieves the highest value in Kappa statistic. It is very interesting to perform a comparison of these results with the performance measures achieved when the full feature space was considered for training (see Table 3). We observe that in general the performance decreases when the feature dimensionality is reduced. But this is not the case for logistic regression. This technique outperforms its full feature space version by 5 accuracy percentage points, increasing also the Kappa statistic value. This improvement does not affect error measures. On the other hand, Bayes networks and random forest decrease accuracy results and increase error measures. From this point of view, the logistic regression-based classifier

achieves the best balance between performance and dimensionality reduction.

We analyze also detailed result per class for the best three learning algorithms. These results are shown in table 6.

Table 6: Detailed results for different learning algorithms over newsworthy detection using feature selection.

			Bayes Net		
	FP Rate	Precision	Recall	F-Measure	ROC Area
NEWS	0.149	0.783	0.637	0.702	0.825
CHAT	0.363	0.735	0.851	0.789	0.825
Weighted Avg.	0.265	0.757	0.753	0.749	0.825
			Logistic		
	FP Rate	Precision	Recall	F-Measure	ROC Area
NEWS	0.216	0.736	0.717	0.726	0.796
CHAT	0.283	0.766	0.784	0.775	0.796
Weighted Avg.	0.253	0.753	0.753	0.753	0.796
			Random forest		
	FP Rate	Precision	Recall	F-Measure	ROC Area
NEWS	0.261	0.706	0.743	0.724	0.798
CHAT	0.257	0.773	0.739	0.756	0.798
Weighted Avg.	0.259	0.743	0.741	0.741	0.798

Table 6 shows that the best false positive rate for the NEWS class is achieved by the Bayes network-based classifier but it achieves also the worst false positive rate for CHAT detection. Logistic regression achieves a good balance for both classes, because it registers improvements in recall regarding the NEWS class. The best weighted F-measure value is also achieved by logistic regression. A very strong result was achieved by Bayes networks regarding the ROC area measure, but with a low recall rate.

4. Learner selection details for predicting credibility

Table 7 shows how these methods perform in each class. The last row of each evaluation shows the weighted average between both classes. As Table 7 shows, all these methods register a high false positive rate for the class "`CREDIBLE`", which indicate to us that it is very easy to miss-classify not credible cases as credible, lying in a very significant false positive rate. However recall rates are very acceptable for the meta learning and random forest algorithms regarding the class "`CREDIBLE`", indicating to us that the performance for credibility prediction is acceptable regarding true positive rates. ROC areas are greater than 0.6 for the three classifiers, illustrating the compromise between hit rate and false alarm. Best values for ROC areas were

Table 7: Detailed results for different learning algorithms over credibility assessing.

	Random forest				
	FP Rate	Precision	Recall	F-Measure	ROC Area
CREDIBLE	0.5	0.618	0.724	0.667	0.639
NOT-CREDIBLE	0.276	0.618	0.5	0.553	0.639
Weighted Avg.	0.394	0.618	0.618	0.613	0.639
	Logistic				
	FP Rate	Precision	Recall	F-Measure	ROC Area
CREDIBLE	0.463	0.609	0.645	0.626	0.611
NOT-CREDIBLE	0.355	0.575	0.537	0.555	0.611
Weighted Avg.	0.412	0.593	0.594	0.593	0.611
	Meta learning				
	FP Rate	Precision	Recall	F-Measure	ROC Area
CREDIBLE	0.522	0.612	0.737	0.669	0.607
NOT-CREDIBLE	0.263	0.619	0.478	0.539	0.607
Weighted Avg.	0.4	0.615	0.615	0.608	0.607

achieved by random forest and logistic regression.

5. Detailed analysis of the credibility classifier thresholds

We explore how output scores perform for the logistic regression credibility classifier. In Figure 1 we show histograms for output scores associated to credible and non-credible labels. Figure 1a shows the histogram for hit cases and Figure 1b shows the histogram for error cases.

As Figure 1 shows, a significant fraction of error cases register output scores under 0.6, illustrating the absence of enough evidence to predict the label. On the other hand, a very significant fraction of hit cases register output scores close to 1, illustrating that in these cases the classifier identifies more evidence to predict the label. A fuzzy region is between 0.6 and 0.7 output scores, approximately, where hit and error cases are registered. A delicate balance between false negatives and false positives is presented in this region, where the inclusion of a score threshold affects this balance.

We study how the inclusion of a score threshold helps discarding cases where the output score was very close to 0.5. Intuitively, a higher score threshold tends to discard more error cases but introducing more false negatives. This situation is illustrated in Figure 2.

As Figure 2 shows, a higher threshold leads in a low coverage rate (see the curve labeled as "filtered"). For example, when we use a threshold equals to 0.6, almost 40% of the predictions are discarded. The precision measure increases from 66.6% to 77.1% in this evaluation. The false-positive rate decreases from 40% to

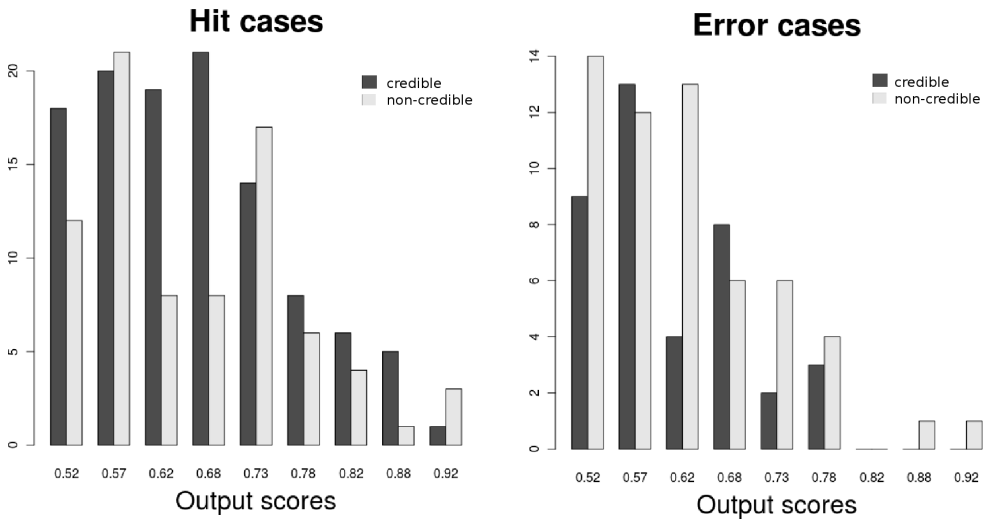

Fig. 1: Output scores for hit and error cases. Black bars are credible labels and gray bars are non-credible labels.

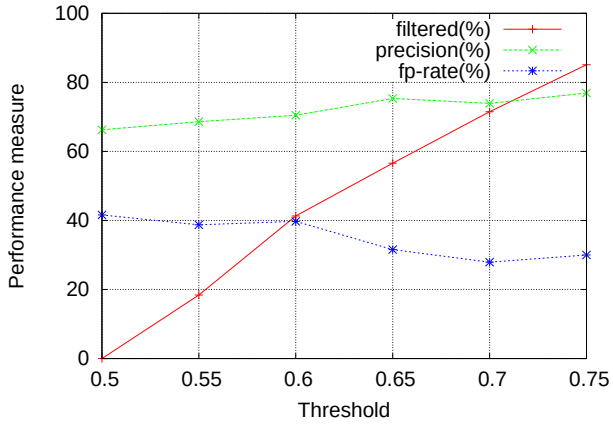

Fig. 2: Precision and false positive rates using an output score threshold.

25%, approximately. An interesting point is registered when the threshold is equals to 0.6, where the 40% of the predictions are filtered and a 70.4% of precision is achieved. Notice that the false positive rate does not reach a significant decrease in this evaluation.

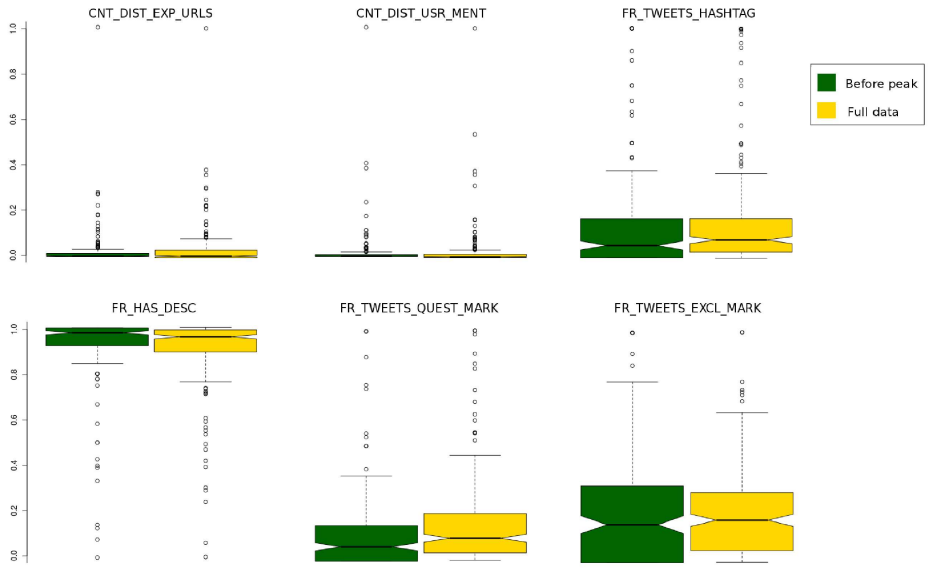

Fig. 3: Boxplots for features with different distributions before the first peak of activity and the full event timeline.

6. Detailed analysis of early prediction classifiers

To compare data we conduct a Kolmogorov-Smirnov test for each pair of distributions. The Kolmogorov-Smirnov test quantifies a distance between two data samples. The null hypothesis of the test is that the samples are drawn from the same distribution. The null hypothesis is rejected if the distance between both samples is significant, considering a given p-value for the significance of the test. Accordingly, the null hypothesis is rejected when the p-value is less than the significance level, which we set as 0.01 (null hypothesis rejected at the 1% significance level). We conduct these tests over the set of features that our classifiers use, considering the eight features used for newsworthy detection and the twenty features used for credibility assessment, that in practice are twenty four features because four features of the newsworthy model are also use by the credibility model. Our results show that the underlying distribution differs significantly for six features, four of them used by the newsworthy model and three used by the credibility model, where one feature is used in both models. We show the boxplots of these features in Figure 3.

Table 8 shows that the early detection of newsworthy topics is feasible. In fact, a false positive rate equals to 36% is registered in the NEWS class, whilst the false positive rate for the CHAT class achieves a 17% of the cases. This indicate to us that several cases initially labeled as CHAT in fact are related to a newsworthy event. However, cases labeled as NEWS are in general related to newsworthy topics, being

Table 8: Detailed results per class over early newsworthy prediction.

	FP Rate	Precision	Recall	F-Measure	ROC Area
NEWS	0.369	0.631	0.824	0.714	0.778
CHAT	0.176	0.824	0.631	0.714	0.778
Weighted Avg.	0.26	0.74	0.714	0.714	0.778

in this case only the 17% of the NEWS cases misclassified as CHAT, suggesting that early newsworthy labels can be used to determine good candidates for news tracking (note that the recall rate if very significant for the NEWS class). Eventually, a second labeling process can be conducted over CHAT cases, trying to detect NEWS cases. The use of more tweets for feature estimation can be useful for this goal, reducing the false positive rate. Finally, note that a ROC area equals to 0.778 is achieved by both classes and that the F-measure is the same for both classes.

Table 9: Detailed results for two 2 credibility classification strategies for early prediction.

			Logistic		
	FP Rate	Precision	Recall	F-Measure	ROC Area
CREDIBLE	0.317	0.804	0.788	0.796	0.86
NOT-CREDIBLE	0.212	0.662	0.683	0.672	0.86
Weighted Avg.	0.278	0.75	0.749	0.749	0.86
			Logistic (Th = 0.6)		
	FP Rate	Precision	Recall	F-Measure	ROC Area
CREDIBLE	0.158	0.81	0.751	0.779	0.857
NOT-CREDIBLE	0.249	0.791	0.842	0.816	0.857
Weighted Avg.	0.412	0.593	0.799	0.798	0.857

7. Model transfer details

As Figure 4 shows, these boxplots illustrate some differences in the use of Twitter under a crisis situation and a normal situation. Some very interesting relations arise from this analysis. We can observe that during the Chilean earthquake posts were short, a significant fraction of them used hashtags and emoticons (in particular the frown emoticon) and RT trees tended to register greater depths.

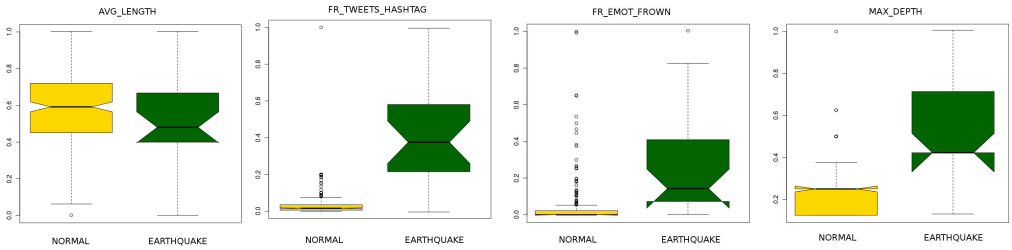

Fig. 4: Boxplots for each of the 4 features of newsworthy detection which register significant differences between a normal situation and the Chilean earthquake.