Electoral Contention and Violence (ECAV) Data Project
Assessment of the coders' identification performances

Elio Amicarelli

doc ref: ECAV09016

1. *Introduction*

Before coding an ECAV event coders have to identify it among all the information contained in a news item. It follows that in order to assess the quality of the ECAV data we have to start from the assessment of our coders' identification performances. This document reports the methodology adopted for this assessment along with results in accordance with the strategy defined in the document reference *ECAV08016*.

2. *Setting and Methods*

Setting

The evaluation setting for identification performances outlined in ECAV08016 involves the following steps:

1. Let S be a set of ECAV articles the coders have never seen before. Create an Identification Gold Standard $IGS = (y_1, y_2, y_3, ..., y_n)$ containing descriptions and general coordinates for all the n ECAV events contained in S . *Status: Completed*
2. Let $C_i = (\hat{y}_{i1}, \hat{y}_{i2}, \hat{y}_{i3}, ..., \hat{y}_{im})$ be the set of m events identified by a general *coder* i over the set S. Obtain C_i for each coder currently hired. *Status: Completed*
3. For each coder establish unique element-wise correspondences between C_i and IGS in order to obtain the set of correctly and mistakenly identified events in C_i (these are respectively the true positives and the false positives). *Status: Completed*
4. Assess coders' identification Accuracy and Precision. *Staus: Finalized in this report*

Methods

In order to assess coders' performances I implemented standard statistical methodologies used for the evaluation of classification systems. More specifically, for each classifier (coder) I examined the classifier's True Positive Rate (TPR) and Precision defined as:

$$TPR = P(C_i = 1|IGS = 1)$$

The TPR is the rate of correct identifications. The TPR ranges from 0 to 1. The TPR equals 1 when a coder correctly identified all the ECAV events of interest; conversely, it equals 0 when a coder failed to identify all the ECAV events of interest.

$$Precision = P(IGS = 1|C_i = 1)$$

Precision is the rate of identifications which turn out to be correct. Precision equals 1 when a coder is correct every time he identifies an event; conversely, it equals 0 when a coder is always wrong when he identifies an event.

For each TPR and Precision value 95% bootstrapped confidence intervals were calculated using the percentile method on 5000 bootstrap samples.

3. *Results*

The anonymized Accuracy and Precision results are showed in Figure 1 while a version containing coders' names can be found in the separate Appendix. The dashed grey line placed at a performance value of 0.5 represents the expected Accuracy performance of a balanced binary classifier producing random guesses.

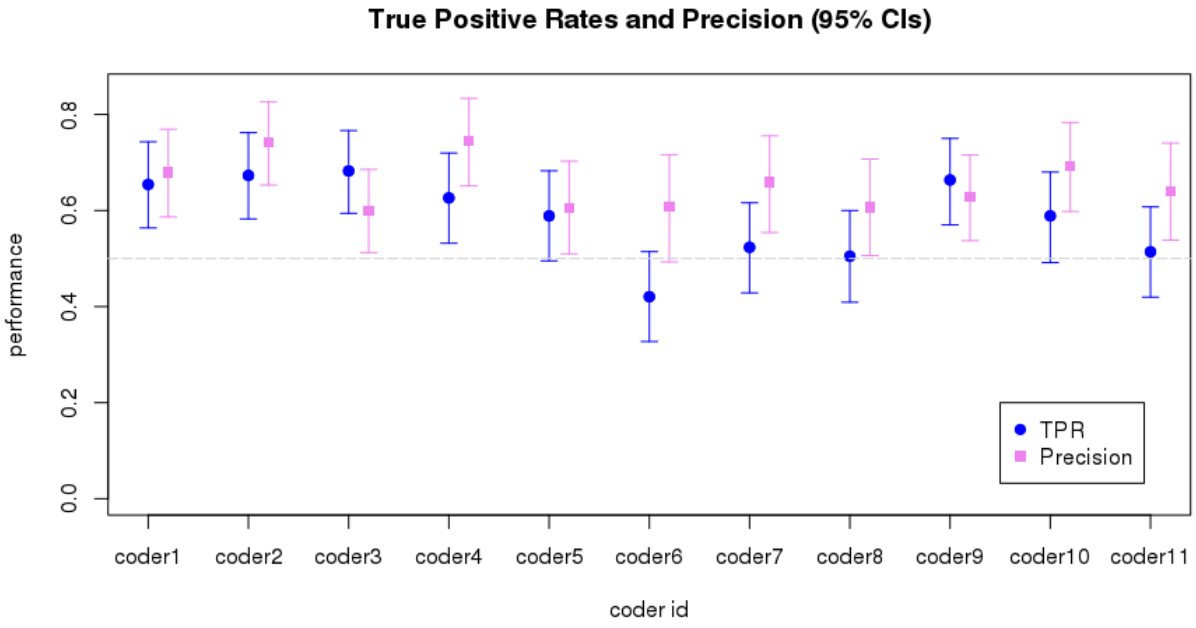

Figure 1: Individual True Positive Rate and Precision for 11 ECAV coders

As showed in Figure 1 coders 1, 2, 3, 4 and 9 have TPRs between 0.6 and 0.68 while coders 5,7,8,10 and 11 are between 0.5 and 0.59. The TPR for coder 6 is 0.42, the lowest in this set. It is worth to remember that a TPR of 0.42 means that coder 6 is correctly identifying only 42% of the actual ECAV events contained in the IGS.

Notice that the random classification threshold (grey line in Figure 1) falls within the TPR confidence intervals for coders 5, 6, 7, 8, 10 and 11. Because of this, we cannot be confident that identification performance of these coders is actually better than a randomized binary classification.

Despite coders 1, 2, 3, 4 and 9 are those doing a better job in identifying ECAV events, the maximum TPR is 68% (coder 3). This suggests that there is room to improve the identification ability of all coders, even those who performed better during this exercise.

Except for coder 3 and coder 9, the point estimates for Precision are always higher than the respective TPR estimates. This suggests that generally coders are more prone to produce False Negatives than False Positives. Because of this, particularly attention should be paid in advising coders 3 and 9 on how to avoid to select irrelevant information,while all the other coders should be advised on how to avoid to overlook relevant information.