Electoral Contention and Violence (ECAV) Data Project Defining a strategy to assess data quality

Elio Amicarelli

doc ref: ECAV08016

1. Background and Objective

The Electoral Contention and Violence (ECAV) Data Project is funded by the Dutch Science Foundation NWO and from a EU Marie Curie Career Integration grant. One of the core objectives of this project consists in extracting political events of interest from a large database of news articles. In order to accomplish this goal, 11 trained human coders read the news articles, *identify* the political events of interest and *encode* each event in machine-readable variables following a set of pre-established coding rules.

Given this setting, several challenges for the quality of the data produced must be considered. Human coders can vary in their ability to correctly identify and encode the political events of interest. There can be several causes for this variation such as *a)* different levels in the understanding of what has to be considered an event of interest, *b)* how the coding rules should be applied in order to translate the news reporting in machine-readable variables and *c)* the effort each coder puts in doing his job. These factors can jeopardize the quality of the data because the level of accuracy of a given event included in the ECAV database would depend in part on the specific characteristics of the coder who produced it.

For these reasons, the Project Leader is interested in assessing the quality of the data produced and in evaluating the coders' performances based on objective criteria. This report outlines an operational strategy to meet these needs via a statistical analysis articulated in 2 stages.

2. Strategy

The process of translating news text into machine-readable event data works as follows: a human coder is assigned a set of news articles, the human coder reads the articles and identifies events that are relevant for the ECAV project, the human coder encodes the relevant events he identified into a set of variables established by the ECAV coding schema.

It is evident that coders' mistakes can happen during two distinct phases:

1. While reading the news the coders may fail to identify relevant events or may mistakenly identify irrelevant events
2. While coding the identified events the coders may fail to code them correctly

In order to evaluate the coders' performances during these two phases, the strategy presented in this section is divided in two stages. *Stage 1* concerns the assessment of the coders' abilities in identifying the event of interest from the news articles. *Stage 2* examines the coders' abilities to encode the event of interest according with the rules established by the ECAV coding schema. This second aspect is examined under two dimensions, which are namely: *a)* The reliability of the data produced and *b)* the validity of the data produced. For a given variable of the ECAV schema, the *reliability* concerns the level of agreement among coders on the values they assigned to it. On the other hand, *validity* measures the level of agreement of each coder with a pre-established correct version of the data. In other words, while reliability gives us information about whether the coders are applying the encoding rules in the same way, validity tells us if the coders are applying the encoding rules correctly.

2.1 *Material Preparation*

Before starting with the two-stages assessment, a correct version of the data to be used as benchmark for the coders' performances must be created. Hereafter I will refer to these data as Identification Gold Standard (the benchmark to be used in Stage 1) and Coding Gold Standard (the benchmark to be used in Stage 2). *The guidelines and modalities for the creation of these two datasets will be discussed direcly with the Project Leader during the next meeting.*

2.2 *Stage 1 - Assessing coders' identification performances*

Description. The aim of this stage is to assess the ability of each coder to correctly identify the events of interest based on the project's identification rules.

Setting. In this first stage, coders are provided with a common set of articles and are asked to identify what they think are relevant ECAV events. They are also provided with a spreadsheet where for each identified ECAV event they have to record *a)* the relevant text snippet containing the event and *b)* the title, date and unique identifier of the article containing the text snippet. In order to do not bias their performances, the coders are told that the aim of the exercise is to cross-check some dubious results produced by previous employees.

Analysis. The events identified by each coder are matched with the events identified and included in the Identification Gold Standard. Based on this last step, a binary vector containing a set of True Positives (coded as 1s) and False Negatives (coded as 0s) is derived for each coder. This binary information is used to calculate the following identification performance measures for each coder:

- True Positive Rate (TPR): 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.
- Positive Predicted Value (PPV): The rate of identifications which turn out to be correct. The PPV 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. PPV is also called Precision.

These performance measures will be augmented with relevant measures of statistical uncertainty. Since the coders are asked to identify events from the same set of articles, and since none of them has read these articles before, this analysis allows to fairly compare their identification abilities.

Output: A technical report detailing and comparing individual coders' identification performances in accordance with the aforementioned analysis.

2.3 *Stage 2 - Assessing coding reliability and validity*

Description. The aim of this stage is to assess the degree of agreement among coders (intercoder reliability) and the ability of each coder to correctly encode the relevant information in machine-readable variables following the project's coding rules (coding validity).

Setting. The coders are provided with a common set of events' descriptions and are asked encode the relevant pieces of information in accordance with the ECAV data schema. As before, in order to do not bias their performances, the coders are told that the aim of the exercise is to cross-check some dubious results produced by previous employees.

Analysis. The intercoder reliability and the coding validity are assessed using the same set of metrics. At the analysis stage the difference between reliability and validity resides in the fact that reliability is evaluated by making comparisons among coders, while validity is assessed by comparing each coder's output against the coded Gold Standard. The comparison is performed using relevant metrics such as generalized kappa measures.

Output: A technical report detailing and describing both the overall degree of inter-coder reliability and the coders' individual validity performances in accordance with the aforementioned analysis.

Conclusion

This document outlined a statistical strategy to assess the coders' performances and the data quality for the ECAV project. Since the data can be affected by two main sources of error, namely failure in event identification and failure in event coding, the evaluation strategy here proposed breaks the assessment in two distinct stages. Upon approval of the current plan, the next step will consist in preparing the needed material as described in section 2.1.