Automated Document Classification for the social sciencesusing supervised machine learning: Trade-offs between reliability and efficiency

Hans Then, Pythea, the Netherlands
Gerard Breeman, Wageningen University, the Netherlands

Paper abstract for the ATISR conference:
February 10-12, 2012,
Taipei, Taiwan

The increasing amount of electronic resources that has become available since 1995 through the Internet provides many opportunities for the social sciences. Automated topic classification systems can help social scientists, politicians, trend-watchers, spin-doctors, and think-tanks to monitor popular trends. Key words typed in internet search engines, and topics that are being discussed in news bulletinsmay be pointers to public, economic or political trends (Jones and Baumgartner 2005).

The increasing amount of electronic information and the complexity of the questions scholars pose to datasets, requires efficient and effective issue classification methods.Many social scientists use manual coding, such as the congressional bills project (www.congressionalbills.org) and the policy agenda-setting project (see www.policyagendas.org). These projects share a policy topic code book, containing 225 policy topics, divided in 9 main topics. They coded a great deal of different policy statements, such as, the State of the Union speeches, Bills, New york Times newspaper articles, executive speeches,European Union council conclusions etcetera. (Adler and Wilkerson 2008; Alexandrovaet al. 2011)

However, manual coding is time consuming and previous research explored the feasibility of supplementing the manual coding with machine learning coding techniques (Hillard et al. 2008; Purpura et al. 2006).Machine coding uses manually coded texts, to inform algorithms about typical features of texts that belong to the various categories of the category system. The goal is to increase the efficiency of manual coding without losing reliability. The conclusion of this research is that the most promising way of coding is to design a supervised machine learning (SML) protocols that combine manual and automated machine coding in an integrated process of learning and training loops (Van Atteveldt et al. 2008, Breeman et al. 2009)

Testing machine learning for the social sciences

In previous research we explored different ways to improve the reliability ofmachine learning software (Breeman et al. 2009). The software package combined 4 to 6 off the shelf algorithms: Support Vector machine, Ling-pipe (2 types), Maxent (2 types), and Naïve Bayes. We used half of the datasets to inform the algorithms and the other half as a test set. The software was tested on the annual executive speech of the Dutch government between 1945-2008, which were topic-coded per sentence (n=8122), the European directives between 1978-2009, coded per Directive (2300), the Dutch coalition agreements between 1963 and 2006, coded per paragraph (n=4455), and Dutch Bills between 1990-2006, coded per Bill (n=2800). The following improvement were tested, against a blank split-half test:

- Word-stemming and stop-word removal: improved results up to 4.5%, depending per algorithm.
- Increasing the size of the training set: on average 1.8% improvement per 500 new manually coded entries.
- Balancing the training set (some of the issue are overrepresented in the trainings sets, the Annual Executive speech eg. contains relatively many statement on foreign affairs issues): this deliveredmixed results. Depending on the type of data-set we found improvements of -2% to 9%.

Thus far, we reached a reliability of up to 70%

Testing SML-protocols

In new test runs we want to introduce a SML-protocol in which the coding-training-testing loop, will be added with a phase where human annotators improve the training-set manually. This is based on the reliability scores that all algorithms produce. After the testing-phase we will take out the 5% entries with the lowest accuracy, manually code/check these entries, and add them to the trainings-set.In addition we also will split up the training and testing sequence into the 19 main topics, instead of training and coding the entire dataset at once.

Trade-off between reliability and efficiency

Furthermore, the paper analyses the trade-off between coding reliability and efficiency. The SML-protocols are meant to improve the efficiency. However, with every extra phase in the protocol that should improve the reliability of the coding, the efficiency will go down; it simply takes more time and effort to code. The question is whether it is always necessary to improve the reliability.

For instance, the goal of the agenda setting project is to single out trends in public, political, and media attention. Wolfe et al. (2009) describe some trends in the New York Times using the manual coding protocol (period 1998-2005, N=22500). We will duplicate this test by using the SML-protocol. Different runs will be conducted in which we use different sizes of the initial training set and different accuracy rates.

We conclude the paper by providing an optimal SML-protocol, in which we find a balance between efficiency and reliability.

References

Alexandrova, P., M. Carammia, and A. Timmermans(2011) *Policy Punctuations and Issue Diversity on the European Council Agenda.*Paper presented at the annual conference of the Comparative Agendas Project, Catania, 23-25 June 2011

Adler, E.S. and J. Wilkerson (2008) ‘Intended consequences? Committee reform and Jurisdictional Change in the House of Representatives’. *Legislative Studies Quarterly*, 33 (1), 85-112.

Atteveldt, W.van, J. Kleinnijenhuis, N. Ruigrok, and S. Schlobach (2008). ‘Good news or bad news: Conducting sentiment analysis on Dutch texts to distinguish between positive and negative relations’ *Journal of Information Technology & Politic*s, 5(1), 73-94.

Breeman, G., H. Then, J. Kleinnijenhuis, W. van Atteveldt, and A. Timmermans (2009) *Strategies for Improving Semi-automated Topic Classification of Media and Parliamentary Documents.* Paper presented at the annual meeting of the Midwest political Science Association, Chicago, April 2-5 2009

Hillard, D., S. Purpura, and J. Wilkerson (2007) ‘Computer-Assisted Topic Classification for the Mixed-Methods Social Science Research’ *Journal of Information Technology and Politics,* Vol. 4(4) 31-46.

Jones, B.D. and F.R. Baumgartner (2005) *The Politics of Attention: how Government Prioritizes Problems.* Chicago: University of Chicago Press.

Purpora, S., and D. Hillard (2006) Automated classification of congressional legislation. In: *Proceedings of the International Conference on Digital Government Research* (pp 219-225), New York: Association for Computer Machines.

Wolfe, M., A.E. Boydstun, and F.R. Baumgartner (2009) *Comparing the Topics of Front-Page and Full Paper Stories in the New York Times.* Paper presented at the annual meeting of the Midwest political Science Association, Chicago, April 2-5 2009.