IDEOLOGICAL SCALING OF SOCIAL MEDIA USERS: A DYNAMIC LEXICON APPROACH

SUPPLEMENTARY MATERIALS

Mickael Temporão[1], Corentin Vande Kerckhove[2], Clifton van der Linden[3], Yannick Dufresne[1], and Julien M. Hendrickx[2]

[1]Université Laval
[2]Université Catholique de Louvain
[3]University of Toronto

1 Generating Network Ideologies

The network estimates are computed from the network model introduced by Barberá (2015), using a No-U-Turn sampler (Hoffman and Gelman, 2014). The network estimates are obtained by averaging 800 generated samples. The sampler requires that priors are determined and initial values set for the model parameters. We consider normal priors on the model parameters as described in Barberá (2015). Elites' ideologies are initialized by performing a unidimensional correspondence analysis on the binary matrix Y. All the other parameters are also randomly initialized as described by Barberá (2015).

mickael.temporao.1@ulaval.ca

The samples are generated with 200 warm-up iterations followed by 1000 iterations and a thinning of 2 on two different chains.

2 N-gram Optimization

We evaluated the performance of three N-gram sizes: unigrams ($N = 1$), bigrams ($N = 2$), and trigrams ($N = 3$). Each N-gram size is evaluated according to three criteria : *quality*, *classification improvement*, and *scope of information*. The case $N = 1$ corresponds to the classical approach designed for manifestos.

The *quality* criterion measures the convergent validity of the textual estimates compared to other established ideological estimates (see Adcock, 2001). Ideally, we would compare the text-based ideology estimates to a reference ideology estimate such as, for instance, one derived from roll call or from survey data. However, strict party discipline blurs the positions of individual candidates by drawing them toward the established position of their party. To test the convergent validity of the textual estimates, we therefore rely on the network estimates for politicians (Barberá, 2015). The *quality* is defined as the Pearson correlation coefficient (ρ) between the vector of estimates derived from social media textual data and the vector of estimates derived from network data (Barberá, 2015; Bond and Messing, 2015). Its value range from 0% (poor quality) to 100% (high quality).

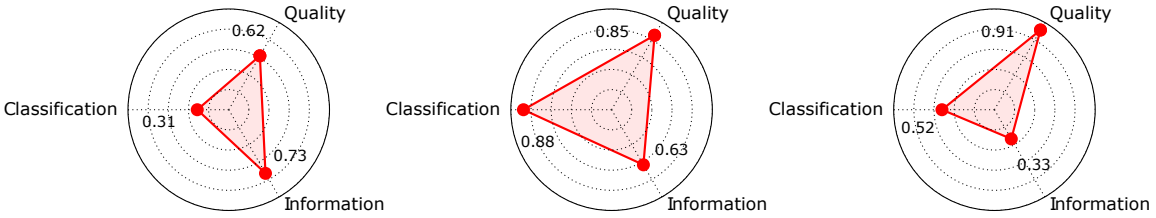
The *classification improvement*, denoted by C_{imp}, measures the complementarity between textual and network estimates. It explains how well the textual estimates improve the detection of the party clusters compared to a simple network classification. This improvement is computed as follows:

$$C_{imp} = \frac{(a_{net+txt} - a_{net})}{(1 - a_{net})} \times 100\%$$

where $a_{net+txt}$ denotes the classification accuracy obtained by using both network and textual estimates, while a_{net} denotes the accuracy achieved by using only the network estimates. For the particular value $C_{imp} = 0\%$, no improvement is made by adding the textual features into the classifier. The elites are perfectly classified ($a = 1$) with both features when $C_{imp} = 100\%$.

The *scope of information* records the percentage of the original sample for which the textual method can provide an estimate. As the value of N increases (i.d., from unigrams to trigrams), the term-document-matrix becomes sparser. This directly reduces the sample of elites for which an estimate can be provided. This is what we call the *scope of information* problem. It is explained by the required filtering constraints on the amount of shared N-grams (see Section **??**). The original sample includes elites for which a network estimation is available.

Figure 1: **Elites - Comparing the performance of the dynamic lexicon approach for unigrams, bigrams and trigrams dictionaries**. Values are obtained by averaging over the three elections. *Information* corresponds to the percentage of the remaining sample after the filtering process. *Quality* corresponds to the convergent validity of textual ideology estimates by performing a Pearson correlation with network ideology estimates. *Classification* corresponds to the average improvement of the classification accuracy between the textual-network classifier and the network classifier. All of the indicators are computed on the shared sample imposed by the trigram filtering conditions.



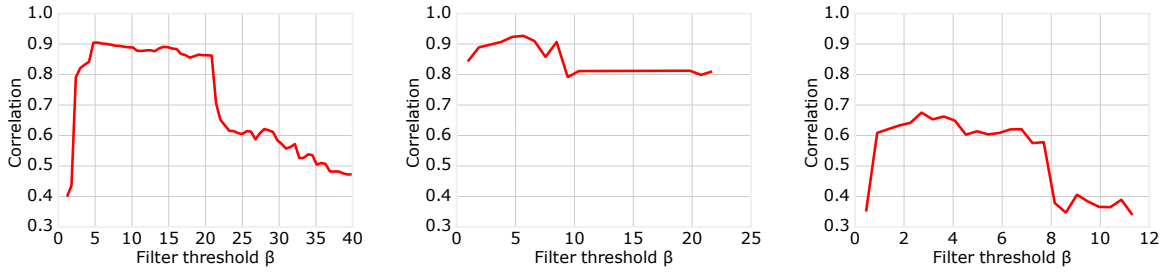
(a) Unigram (b) Bigram (c) Trigram

Overall, Figure 1 shows that, with the increase of N-grams, the *quality* also increases, but this results in a substantive loss of information. In terms of classification, bigrams seem to perform better than single words and trigrams. Overall, working with bigrams faces the entails a trade-off between validity and sample size. It outperforms other N-grams at classification while maintaining reasonable *scope of information* and *quality*.

In addition, we examine the effects of the filtering conditions on the *quality* of the textual estimates for a bigram ($N = 2$) dictionary. Low filters ($\beta < 3\%$) tend to keep noisy n-grams and higher filters ($\beta > 15\%$) tend to suppress relevant information for the

positioning of political candidates. Consequently, the ideal dictionary should take into account this filtering trade-off. These effects are shown in Figure 2.

Figure 2: **Validity of the dynamic lexicon estimates compared to filtering conditions** The x-value represents the filter threshold β. The y-value represents the Pearson correlation (ρ) between the network and textual estimates for the political candidates.



(a) Canada 2015 (b) New Zealand 2014 (c) Quebec 2014

The above leads us to fix the dictionary parameters at $N = 2$ (bigrams) and $\beta = 5\%$ throughout this paper.

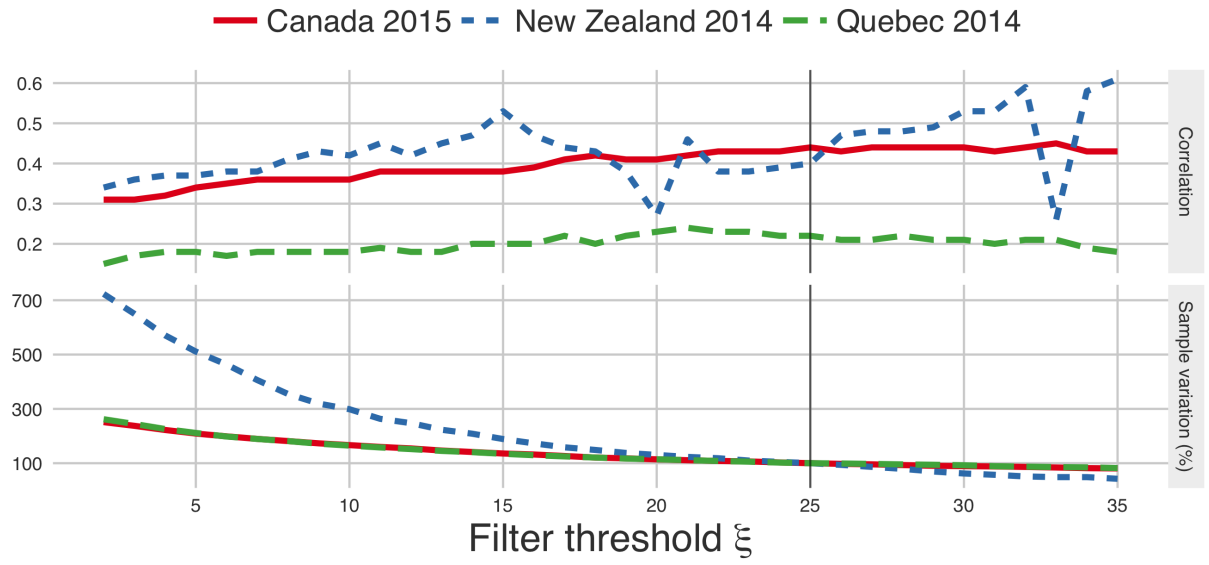
3 Citizens' Bigram Threshold Filtering

Figure 3 illustrates how the citizens' bigram filtering criteria (ξ) affects the quality of the estimated ideologies and the relative sample size. The bigram filtering criteria (ξ) depicts the minimum number of political bigrams in the dynamic lexicon required for citizens to be kept in the sample. The first graph in Figure 3 compares the Pearson correlations (ρ) between the textual estimates and the reference survey ideology estimates for $= 2, ..., 35$. The second graph in Figure 3 illustrates for $= 2, ..., 35$ the evolution of the sample size relative to the predefined sample ($\xi = 25$) in Section 3.

We can see on Figure 3 that as one decreases –that is, when we relax the number of political terms required by citizens–this allows an increase in sample size at the cost of a linear decrease in the quality of the estimate. The increase in sample size is the especially striking for New Zealand 2014. This is partly due to the smaller sample that we had

access to in the New Zealand 2014context. This also suggests that further research should explore the optimization of this filtering criteria (), overall or within each context, as a function of the quality of the estimates and the size of the sample.

Figure 3: **Evolution of the quality of citizens' estimates compared to the bigram threshold filter (ξ).** The x-value represents the filter threshold ξ; that is, the minimum number of political terms required by a given citizen to be kept in the sample. In the first graph, the y-value represents the Pearson correlations (ρ) between the textual ideology estimates and reference survey ideology estimates for the citizens. The second graph's y-value represents the evolution of the sample size as a proportion (%) relative to the predefined sample size ($\xi = 25$) in Section 3.



4 Machine Learning Classification Task Setup

The textual and network ideologies, as well as the survey ideologies for citizens, are treated as features or predictors while the parties are treated as categories or labels. Predicting affiliation and intentions can be formulated as a multi-label classification task since more than two parties are generally involved. We deal with the multiple labels by introducing successive and independent binary classifiers (Read et al., 2011). Each party is associated with one single classifier. In the case of citizens, the classifier predicts

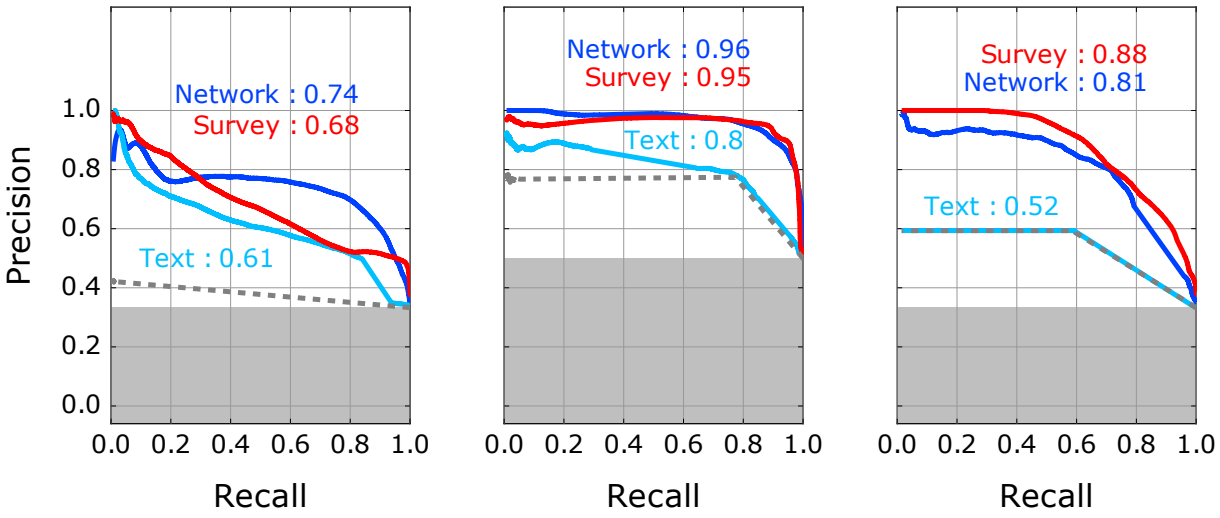
for every citizen whether or not they intend to vote for a given party. For political elites, the decision boundary attempts to separate the candidates belonging to a specific party from the rest of the elites. Classification indicators such as accuracy, precision, and recall are computed for each party. Only individuals that reported their voting intention in the survey are kept in the classification analysis. Also, parties with a very low number of voters are discarded to prevent sampling issues and highly imbalanced classes.

The use of *Support-Vector Machine* classifiers with Gaussian Kernels is recommended to handle a smaller number of features (here, from 1 to 3) and medium size training samples (Hsu et al., 2003). We prevent over-fitting by following a three-fold cross-validation process to train the SVM classifier. That is, the training of the classifier is performed on a dataset including $2/3$ of the sample (training set) and the prediction's performance is assessed on the remaining $1/3$ of the sample (cross-validation set). The performance of one feature is assessed by micro-averaging the performance indicators assessed on the cross-validation set. Our classifier requires us to determine two parameter values: the variable γ composing the Gaussian Kernel $K(x_i, x_j) = \exp(-\gamma(||x_i - x_j||^2)$, and the value of C which controls the cost of misclassified training samples (Smola and Vishwanathan, 2008). We use the arbitrary values of $\gamma = 1/2$ and $C = 1$ in the present analysis. Using an SVM classifier requires us also to determine a threshold value on the decision boundary. Applying different thresholds generates multiple pairs of precision and recall, leading to a precision-recall curve for each party.

Accuracy is the fraction of all instances that are correctly categorized. Precision is also known as positive predictive value and measures the number of selected items that are correctly classified. Recall is also known as the true positive rate and measures the number of correctly classified items that are selected.

Micro-averaging is performed by averaging the curves in proportion to the number of voters for each party.

Figure 4: **Citizens - Average efficiencies of citizens' ideologies to predict vote intention.** Each curve displays the micro-averaged precision and recall obtained by varying the threshold of a Support Vector Machine classifier with a Gaussian Kernel. The classifier is trained using 2/3 of the citizens' ideologies and the precision-recall couple is computed on a 1/3 validation set. The grey area serves as a lower bound and is derived from a random classifier. The grey dotted line represents a baseline when the classifier is trained on the labels' distribution. Values above the graph represent the area under each curve (AUC).



(a) Canada 2015 (b) New Zealand 2014 (c) Quebec 2014

The classification efficiency of a particular feature is assessed by micro-averaging the precision-recall curves for each election, which is illustrated in Figure 4. These are compared to two classical baselines: the average precision of the random classification of the members (grey area), which corresponds to 1 over the number of parties, and the precision obtained by training the SVM on the label distribution (grey dotted line). The average performances of the features are compared by considering the *Are Under the Curve* (AUC) measure, obtained by integrating the precision function over the recall interval. We observe that the *network* and *survey* based estimates perform similarly, from intermediate precision (AUC $>$ 0.65 for Canada) to high precision (AUC $>$ 0.80 for Québec and New Zealand). These two features outperform the *text* based approach, which demonstrates little if any improvement compared to the label-trained baseline.

References

Adcock, Robert. 2001. "Measurement validity: A shared standard for qualitative and quantitative research." *American Political Science Association* 95(03):529–546.

Barberá, Pablo. 2015. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data." *Political Analysis* 23(1):76–91.

Bond, Robert and Solomon Messing. 2015. "Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook." *American Political Science Review* 109(01):62–78.

Hoffman, Matthew D and Andrew Gelman. 2014. "The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research* 15(1):1593–1623.

Hsu, Chih-Wei, Chih-Chung Chang, Chih-Jen Lin et al. 2003. "A practical guide to support vector classification.".

Read, Jesse, Bernhard Pfahringer, Geoff Holmes and Eibe Frank. 2011. "Classifier chains for multi-label classification." *Machine learning* 85(3):333–359.

Smola, Alex and SVN Vishwanathan. 2008. "Introduction to machine learning." *Cambridge University, UK* 32:34.