Automatic classification of open-ended questions: check-all-that-apply questions

Schonlau, Matthias, University of Waterloo, Canada

Gweon, Hyukjun, Western University, Canada

Wenemark, Marika, Linköping University, Sweden and Region Östergötland, Sweden

Abstract

Text data from open-ended questions in surveys are challenging to analyze and are often ignored. Open-ended questions are important though because they do not constrain respondents' answers. Where open-ended questions are necessary, often human coders manually code answers. When data sets are large, it is impractical or too costly to manually code all answer texts. Instead, text answers can be converted into numerical variables and a statistical / machine learning algorithm can be trained on a subset of manually coded data. This statistical model is then used to predict the codes of the remainder.

We consider open-ended questions where the answers are coded into multiple labels (all-that-apply questions). For example, in the open-ended question in our Happy example respondents are explicitly told they may list *multiple* things that make them happy. Algorithms for multi-label data take into account the correlation among the answer codes and may therefore give better prediction results. For example, when giving examples of civil disobedience, respondents talking about "minor non-violent offenses" were also likely to talk about "crimes". We compare the performance of two different multi-label algorithms (RAKEL, CC) to the default method of binary relevance (BR) which applies single-label algorithms to each

code separately. Performance is evaluated on data from three open-ended questions (Happy, Civil Disobedience, and Immigrant).

We found weak bivariate label correlations in the Happy data (90^{th} percentile: 7.6%), and stronger bivariate label correlations in the Civil Disobedience (90^{th} percentile: 17.2%) and Immigrant (90^{th} percentile: 19.2%) data. For the data with stronger correlations we found both multi-label methods performed substantially better than BR using 0/1 loss ("at least one label is incorrect") and had little effect when using Hamming loss (average error). For data with weak label correlations, we found no difference in performance between multi-label methods and BR.

We conclude that automatic classification of open-ended questions that allow multiple answers may benefit from using multi-label algorithms for 0/1 loss. The degree of correlations among the labels may be a useful prognostic tool.

Keywords:

Open-ended questions, multi-label, check-all-that-apply, machine learning, statistical learning, text

Introduction

Text data from open-ended questions can be a valuable supplement to closed-ended questions in surveys. Open-ended questions are important because they do not constrain respondents' answers and may improve the respondent's possibilities to be heard and give

accurate information. They may also provide deeper understanding and insights to the researcher. Included among our examples is a special type of open-ended questions called "probing questions". Probing questions have been shown to be useful as a supplemental technique to cognitive interviewing (Behr, Kaczmirek, Bandilla, & Braun, 2012) and for testing item validity (Behr, Braun, Kaczmirek, & Bandilla, 2013). Although a small number of interviews can help detect major problems with items, larger sample sizes have the advantage to allow quantifying and analysing the findings. Researchers are also starting to collect text answers through recording respondents voice (Revilla, Couper, & Ochoa, 2018), a further use of open-ended questions. Such audio data can be automatically translated into text data (albeit with some introduced error). Survey researchers have also started to collect text data from interactive text messages (Schober et al., 2015) and are paying increasing attention to how open-ended questions are best asked on mobile devices (Peytchev & Hill, 2010; Revilla & Ochoa, 2016).

Text data from web surveys or audio data have increased the interest of using open-ended questions since the respondents' write or speak their answers and no extra annotation work for the researcher is required to create the data set. Coding and analysis of open-ended answers remain however a challenge as they often involve human coders to manually code answers. When data sets are large, it is impractical or too costly to manually code all answer texts. In such cases it is valuable to train a statistical learning model on a subset of the data – the so-called training data- and to use this model to code the remainder of the data. Before the statistical learning model can be employed, text answers are converted into numerical variables, which is explained further below. Such models have been tried for open-ended

questions that generate one answer such as employment (Gweon, Schonlau, Kaczmirek, Blohm, & Steiner, 2017; Schierholz, 2014; Schierholz, Gensicke, Tschersich, & Kreuter, 2018). This situation is comparable to a multiple choice question where one of multiple categories must be chosen (except the categories to choose from are not presented to the respondent).

A further challenge in coding and analysing results from open-ended questions is that respondents often express several points. They may even be prompted to do so. This will generate data comparable to a check-all-that-apply question which allow respondents to make multiple choices. In this paper, we consider open-ended questions where multiple codes can be applied to a respondent's response. Each code assigned represents the presence of a different idea expressed in the response. To automate the tedious and expensive process of coding responses (which is worsened by the possibility of multiple codes per response), we apply statistical learning algorithms for classification. In this setting, each possible code represents a label that is predicted for a given response. Algorithms for multi-label data take into account the correlation among the answer codes and therefore may give better prediction results. To our knowledge, there is no published research on the use of multi-label algorithms to code answers for all-that-apply open-ended survey questions.

We compare the performance of three different multi-label algorithms of which two take into account the correlation between labels. Performance is measured in terms of how well the three methods predict the classification of the manual codes in three different data sets.

Turning text data into n-gram variables

Answers to open-ended questions are text data. A common approach to analysing text data is to create ngram variables based on the text. Each n-gram variable indicates the presence or absence of a single word (“unigram”) or a short word sequence (“bigrams”, “trigrams”, etc.) in the text. Instead of indicator variables for presence or absence, counts can be employed, but are usually unnecessary. This creates typically several thousand n-gram variables. To reduce the number of n-gram variables created and to make them more relevant a number of modifications are usually needed: removing stop words (frequent words like “the” that are not interesting for prediction), stemming (chopping off the tail end of a word so that “walking” and “walks” lead to the same stem), and thresholding (ignoring n-grams that occur less than, say, 5 times across all texts).

For example, Table 1 shows how a simple text has been converted into n-gram variables (unigrams and bigrams) after removing stopwords (“I”, ”am”, ”to”, ”the”) and after stemming (on “emergency” and “going”). The number of words of the text is an additional variable that is frequently useful for prediction.

[Table 1 approximately here]

This technique can be easily employed for Western languages. More details on the n-gram approach to text mining can be found in computer science books (Büttcher, Clarke, & Cormack, 2010, chapter 3) and are also described in Schonlau, Guenther, and Sucholutsky (2017).

For classification, the open-ended question label (based on a manual coding) is then regressed on the n-gram variables. Because most words do not appear in most texts, the design or X-matrix is very sparse, i.e. has a lot of zeroes. Additionally, the large number of variables may exceed the number of observations. This makes it difficult to use logistic regression, and instead machine learning models such as support vector machines (SVM), boosting, or random forests are used.

The n-gram approach and the statistical learning approaches mentioned are all implemented in Stata by the authors and colleagues (Guenther & Schonlau, 2016; Schonlau, 2005; Schonlau et al., 2017) (also "net search randomforest" in Stata for randomforests).

Classifying multi-label text data using multi-label algorithms

The most common approach to classifying multi-label data is binary relevance (BR) (Tsoumakas & Katakis, 2007). Assume there are L labels. Each label is an indicator variable of whether or not the label is present or absent. BR treats the multi-label problem as L separate single-label regression problems. This approach ignores the correlation among the labels. It is the obvious strategy when the researcher does not have access to an implementation of a multi-label algorithm.

At the other extreme, the Label Powerset (LP) (Tsoumakas & Katakis, 2007) method treats each labelset as a single label. In a dataset with e.g. 13 binary labels LP will transform the multi-label problem into a single-label problem with $2^{13} = 8192$ labels (not all of which may occur in the data). While LP does take into account correlations, the large number of classes renders this method useless for practical purposes for most applications.

The random k-labelsets method, RAKEL (Tsoumakas, Katakis, & Vlahavas, 2011; Tsoumakas & Vlahavas, 2007), is a variation on the LP method. Instead of considering all L labels, RAKEL chooses a subset of k labels at random and predicts just this subset of labels. This is repeated m times, each time choosing a new random subset of k labels. Prediction is then made by majority vote, i.e. the modal class (the class most frequently predicted among the m iterations) is predicted. The authors of this method recommend using k=3 and m=2*L tries where L is the number of labels.

Classifier chains (CC) are an extension of BR (Read, Pfahringer, Holmes, & Frank, 2011). As in BR, labels are predicted one at a time. However, the regression for the second label includes an additional x-variable: the predicted first label. The regression for the third label includes two additional x-variables: the predicted first and second labels. This continues until all labels are predicted. This method is affected by the ordering of the labels. To reduce the dependence on the variable order, an ensemble version of classifier chains, ECC, has been proposed by the same authors. ECC runs the CC algorithm repeatedly (e.g. 50 times) as follows: apply CC to a bootstrap sample of the original data using a random label order. For a given text, each of the multiple labels is either predicted to be present or not in each chain (e.g. 50 predictions for each label). The overall prediction could be obtained by rounding the fraction of predicting a given label (e.g. 27 out of 50 times) to either 0 or 1. Rounding corresponds to using a classification threshold of 0.5. Instead, the classification threshold is set such that the average number of labels per text in the test data match that of the training data. Because ECC is superior to CC, typically only ECC results are reported.

Additional methods to multi-label classification are described, for example, in Tsoumakas and Katakis (2007) and Gweon (2017). The purpose of this paper is to show the usefulness of multi-label methods in the context of open-ended questions, rather than an exhaustive comparison of methods. ECC and RAKEL were chosen because they are representative multi-label algorithms (Zhang & Zhou, 2014) with high predictive performance (Madjarov, Kocev, Gjorgjevikj, & Džeroski, 2012).

Data

Performance of classification based on multi-label algorithms is evaluated using 10-fold cross validation on data from three open-ended questions: 1) things that make you happy (in Swedish), 2) give examples of civil disobedience (in German/Spanish/Danish) and 3) meaning of the word immigrant (in German). These three open-ended questions have been coded manually as check-all-that-apply questions since respondents may express more than one answer. Throughout, these three data sets will be referred to as Happy, Immigrant and Civil Disobedience.

To understand how positive factors relate to mental health and care needs for mental health problems (Wenemark et al., 2018) Wenemark et al. (2018) asked respondents "Name some positive things in your life, that are uplifting or make you happy: (you may write several things) ". The answers are classified into 13 codes (labels): nothing, relationships (family or romantic), working/studying, health, self-esteem, joy/happiness, well-being: drinking/eating/drugs/sex, spirituality, money, nature, hobbies, culture, exercise. The question was asked in Swedish. The data set contains n=2350 respondents.

To understand cross-cultural equivalence about civil disobedience, Behr, Braun, Kaczmirek, and Bandilla (2014) first asked a closed item from the ISSP (ISSP Research Group, 2012) “How important is it that citizens may engage in acts of civil disobedience when they oppose government actions?” (Not at all important 1 --- Very important 7), and then probed respondents’ comprehension with an open-ended question: “What ideas do you associate with the phrase `civil disobedience’? Please give examples.”

Answers were classified into 12 codes (labels): non-productive, violence, disturbances, peaceful, listing activities, breadth of actions, breaking law, breaking rules, government dissatisfaction, government-deep rift, copy/paste from the Internet, other. Here we use the combined Spanish, Danish and German data. For the Spanish and Danish data a translation was available into German, so that the analysis was done in German. This may make the classification harder as the translation might have added noise. The merged data set contains n=1029 respondents and can be downloaded from http://dx.doi.org/10.7802/1795.

To study cross-national equivalence of measures of xenophobia, Braun, Behr, and Kaczmirek (2013) classified answers to open-ended questions on beliefs about immigrants. These questions were part of the 2003 International Social Survey Program (ISSP) on National Identity. Here we focus on the German survey with n=1006 respondents. The questionnaire contained four statements about immigrants such as “Immigrants take jobs from people who were born in Germany”. Respondents had to rate this statement on a 5-point Likert scale (“strongly disagree” to “strongly agree”). After each of the four questions respondents were then probed with an open-ended question: “Which type of immigrants were you thinking of when you answered the question? The previous statement was: [text of the corresponding

item].” Answers were classified into 14 codes (labels): non-productive, positive, negative, neutral/work, general, Muslim countries, eastern European, Asia, ex-Yugoslavia, EU15, sub Sahara, Sinti/Roma, legal/illegal, other. The data can be downloaded from http://dx.doi.org/10.7802/1795.

Method

We first explore bivariate correlations among the labels. We then transformed the text into n-gram variables (unigrams only) using language specific stemming and keeping stopwords. Next, we apply multi-label algorithms to the three data sets and compare the predictions to the manual classification in terms of Hamming loss and 0/1 loss. To explain what this means, imagine that we have two respondents whose answers have been classified into 10 labels resulting into a total of 20 classifications. For the first respondent all 10 labels were predicted correctly (10/10 correct) and for the other 9 of 10 labels were predicted correctly (9/10 correct). Hamming loss computes the average error rate of all predicted labels. In our example 19 of the 20 labels are predicted correctly, and the Hamming loss is 1/20=0.05 or 5%. 0/1 loss computes the average number of respondents with at least one incorrect label. In our example one of the two respondents had all 10 labels predicted correctly, and the 0/1 loss is 1/2=0.5 or 50%. Thus, 0/1 loss is a much stricter criterion than Hamming loss. You might prefer 0/1 loss, for example, if you want to manually verify all texts for which at least one label was hard to estimate (low probability).

For each data set a 10-fold cross validation was performed. Cross validation splits data into training data (on which algorithms are build) and test data (on which predictions are done).

The cross validation is done multiple times (here 10 times) in a way such that each respondent is part of the test data exactly once, and then averages the results.

We compare the performance of multi-label algorithms (RAKEL, ECC) with the default strategy, BR. The multi-label algorithms require a base classifier. We use SVM (Vapnik, 2000) with a linear kernel throughout because the classification task often becomes nearly linearly separable for text data due to the large number of features (Joachims, 1998). For the RAKEL algorithm, we used the author's recommendations as described above. The ECC algorithm used 10 bootstrap samples.

Results

Correlation of labels

Figure 1 visualizes the matrix of bivariate correlations for each of the three data sets. Larger circles correspond to stronger correlations. Negative correlations are colored red, positive correlations are colored blue (In the print version the colors are light grey and black). The diagonal corresponds to a correlation of 1 as usual denoted by large blue circles.

[Figure 1 approximately here]

The Happy data have relatively low bivariate correlations overall whereas the Immigrant and Civil Obedience data sets have stronger correlations. For the immigrant data, the correlation matrix plot in Figure 1 shows a blue square (print version: black square) of positively correlated variables. These labels correspond to the naming of country groups (Muslim

countries, Eastern European countries, etc.). The correlation can be explained as follows: If a respondent chooses to answer this question by naming a group of countries, he/she is more likely to name multiple such groups.

The average number of labels per observations and summary statistics about the bivariate correlations are shown in Table 2. There are (L choose 2) bivariate correlation, where L is the number of labels. Multi-label algorithms take into account correlations among the labels; but average correlation may not be the best indicator of how much the multi-label algorithms will benefit. The algorithms may benefit more from a smaller number of larger correlations rather than a larger number of smaller correlation. Therefore, Table 2 shows various percentiles in the right hand tail of the bivariate correlations (75^{th}, 90^{th} 95^{th}).

The Happy data have weaker bivariate correlation than the other two data sets based on average correlation, all percentiles and maximum correlation. However, the Happy data has a much larger average number of labels in the data. The average number of labels is comparable here because the number of labels, L, is similar for the three data sets.

[Table 2 approximately here]

Applying Multi-label methods

The results are shown in Table 3 for 0/1 Loss and Table 4 for Hamming Loss. For 0/1 Loss on the data sets Civil Disobedience and Immigrant, the multi-label methods ECC and RAKEL result in substantially better performance (smaller loss) as compared to BR. ECC is doing somewhat better than RAKEL. For the Happy data there is virtually no difference between BR,

ECC and RAKEL. The multi-label methods perform well for the two data sets with stronger label correlations.

[Table 3 approximately here]

For Hamming loss, there was little difference between RAKEL, ECC and BR on all three data sets. Table 2 shows that the data sets have only 1.15-2.77 labels turned on, on average, and that therefore most of the 12-14 labels in these data sets are zero.

[Table 4 approximately here]

We investigated for which classes the multi-label algorithms ECC/RAKEL improved accuracy most relative to BR. When the true label was "non-productive" in the immigrant data, the accuracy of BR, ECC and RAKEL were 37%, 95% and 92%, respectively. BR often classified as no label (all labels turned off), whereas ECC and RAKEL rarely did. When the true label was "other" in the immigrant data, the accuracy of BR, ECC and RAKEL were 28%, 42% and 32%, respectively. The most common misclassification for BR was again no label at all; whereas ECC and RAKEL rarely classified as no label. In this particular case ECC did better than RAKEL because RAKEL often predicted a second label in addition to "other". When the true label was "government-dissatisfaction" in the civil disobedience data, the accuracy of BR, ECC and RAKEL were 22%, 34% and 30%, respectively. Most of the gains for ECC came from classifying BR's no label to the correct label; the gains for RAKEL did not have a dominant factor. We observed a

similar result for the label “other” in the civil disobedience data where most gains from ECC came from reclassifying no label; and gains for RAKEL were multi-faceted.

Discussion

Multi-label open-ended answers are common and many open ended questions in surveys are asked with the ultimate goal of classifying the answer texts into categories (Wenemark et al., 2018). Classifying answers automatically can save time, resources (in reducing the texts that require manual classification), and make the classification easily repeatable.

We find weak bivariate label correlations in the Happy data, and stronger bivariate label correlations in the Immigrant and Civil Disobedience data. For the data with stronger correlations we found both multi-label methods performed substantially better than BR using 0/1 loss and had little effect when using Hamming loss. For data with weak label correlations, we found little difference in performance between multi-label methods and BR.

Why is the Hamming relatively insensitive to the algorithms used? The Hamming loss refers to the average number of misclassified labels. Most labels are turned off (0) and are always predicted correctly. The prediction of the remaining labels does not affect the *average* label prediction much. The 0/1 loss is more sensitive to misclassification because a mistake in classifying one label counts the whole labelset as incorrect. The 0/1 correlation rewards correctly predicting the co-occurrence among the labels, whereas the Hamming loss – assessing one label prediction at a time- is relatively insensitive to it.

In comparing the relative strength of algorithms, in addition to Hamming loss and 0/1 loss computer scientists consider additional evaluation criteria such as macro F (Read et al., 2011). We chose Hamming loss and 0/1 loss because we believe that the average number of misclassified labels and whether or not the whole labelset is predicted correctly are most relevant to the typical goals of social scientists.

Limitations of the study include the following: First, we have illustrated the multi-label open-ended questions with three data sets. Although the questions are asked in different languages, and consider vastly different topics there is no guarantee that they are representative of all kinds of open ended questions. Second, it is not possible to quantify how large the bivariate correlations have to be before the multi-label algorithms yield a substantial improvement in 0/1 loss. However, for any given data set one can simply split the data into training and test data to explore how much a multi-label algorithm improves 0/1 loss relative to the default algorithm, BR. Third, multi-label algorithms require a base learner for classification. We chose SVMs here. We used Random Forest and Naïve Bayes as base learners and achieved similar results: for the data with stronger correlations multi-label methods performed substantially better than BR using 0/1 loss.

From a practical point of view, some additional questions arise. First, how large should a data set be to make automatic classification more attractive over manual classification? For multi-class problems, we (Schonlau & Couper, 2016) have argued automatic classification becomes attractive when at least about 2,000 observations need to be classified. 500 of these observations need be manually coded for training; this leaves about 1500 observations to be automatically classified. We believe for multi-label problems the same recommendation

applies. Of course, complicated topics that need a large number of labels will require larger training data sets. Further empirical research is needed to explore the tradeoff between the size of the training data and coding accuracy for all-that-apply open-ended questions.

Second, this paper has addressed the question whether correlations among the labels can be exploited to yield better predictions. A related question is what to do if the predictions are not "good enough", on average, as measured by the criterion chosen. In this case we advocate semi-automatic prediction for single label prediction: easy-to-predict text answers are classified automatically, and hard-to-predict text answers are classified manually (Schonlau & Couper, 2016). However, in multi-label classification there is one classification probability associated with each label. What easy-to-classify means for the whole text answer is no longer obvious. A paper about semi-automatic prediction for multi-label data is in preparation.

We conclude that automatic classification of open-ended questions that allow multiple answers may benefit from using multi-label algorithms for 0/1 loss. The degree of correlations among the labels may be a useful prognostic tool.

Acknowledgements

This research was supported by the Social Sciences and Humanities Research Council of Canada (SSHRC # 435-2013-0128). Dr. Dorothée Behr kindly provided us with the GESIS Immigrant data and the GESIS Civil Obedience data. We are grateful for the data.

Matthias Schonlau is Professor in the Department of Statistics at the University of Waterloo, Canada. He can be reached at schonlau@uwaterloo.ca.

Hyukjun Gweon is Assistant Professor of Statistics at Western University, Canada. He can be reached at hgweon@uwo.ca .

Marika Wenemark is Associate Professor in the Department of Medicine and Health Sciences, Division of Community Medicine at Linköping University, Sweden, and Centre for Organisational Support and Development, Region Östergötland, Sweden. She can be reached at marika.wenemark@liu.se.

References

Behr, D., Braun, M., Kaczmirek, L., & Bandilla, W. (2013). Testing the validity of gender ideology items by implementing probing questions in web surveys. *Field Methods, 25*(2), 124-141.

Behr, D., Braun, M., Kaczmirek, L., & Bandilla, W. (2014). Item comparability in cross-national surveys: results from asking probing questions in cross-national web surveys about attitudes towards civil disobedience. *Quality & Quantity, 48*(1), 127-148.

Behr, D., Kaczmirek, L., Bandilla, W., & Braun, M. (2012). Asking Probing Questions in Web Surveys Which Factors have an Impact on the Quality of Responses? *Social Science Computer Review, 30*(4), 487-498.

Braun, M., Behr, D., & Kaczmirek, L. (2013). Assessing cross-national equivalence of measures of xenophobia: Evidence from probing in web surveys. *International Journal of Public Opinion Research, 25*(3), 383-395.

Büttcher, S., Clarke, C. L., & Cormack, G. V. (2010). *Information Retrieval: Implementing and Evaluating Search Engines*. Cambridge, Massachusetts: MIT Press.

Guenther, N., & Schonlau, M. (2016). Support vector machines. *Stata Journal, 16*(4), 917-937.

Gweon, H. (2017). *Statistical Learning Approaches to Some Classification Problems.* (Ph.D.), University of Waterloo.

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., & Steiner, S. (2017). Three methods for occupation coding based on statistical learning. *Journal of Official Statistics, 33*(1), 101-122.

ISSP Research Group. (2012). ZA3950: International Social Survey Programme: Citizenship - ISSP 2004. Version 1.3.0. GESIS Data Archive, Cologne. DOI: 10.4232/1.11372.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), *European conference on machine learning (ECML)* (pp. 137-142). Berlin: Springer.

Madjarov, G., Kocev, D., Gjorgjevikj, D., & Džeroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. *Pattern recognition, 45*(9), 3084-3104.

Peytchev, A., & Hill, C. A. (2010). Experiments in mobile web survey design: Similarities to other modes and unique considerations. *Social Science Computer Review, 28*(3), 319-335.

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning, 85*(3), 333-359.

Revilla, M., Couper, M. P., & Ochoa, C. (2018). Giving Respondents Voice? The Feasibility of Voice Input for Mobile Web Surveys. *Survey Practice, 11*(2), January 22, 1018.

Revilla, M., & Ochoa, C. (2016). Open narrative questions in PC and smartphones: is the device playing a role? *Quality & Quantity, 50*(6), 2495-2513.

Schierholz, M. (2014). *Automating Survey Coding for Occupation.* (Master's Thesis), University of Munich.

Schierholz, M., Gensicke, M., Tschersich, N., & Kreuter, F. (2018). Occupation coding during the interview. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 181*(2), 379-407.

Schober, M. F., Conrad, F. G., Antoun, C., Ehlen, P., Fail, S., Hupp, A. L., . . . Zhang, C. (2015). Precision and disclosure in text and voice interviews on smartphones. *PloS one, 10*(6), e0128337. doi:https://doi.org/10.1371/journal.pone.0128337

Schonlau, M. (2005). Boosted Regression (Boosting): An Introductory Tutorial and a Stata Plugin. *The Stata Journal, 5*(3), 330-354.

Schonlau, M., & Couper, M. P. (2016). Semi-automated categorization of open-ended questions. *Survey Research Methods, 10*(2), 143-152.

Schonlau, M., Guenther, N., & Sucholutsky, I. (2017). Text mining with n-gram variables. *Stata Journal, 17*(4), 866-881.

Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining, 3*(3), 1-13.

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2011). Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering, 23*(7), 1079-1089.

Tsoumakas, G., & Vlahavas, I. (2007). *Random k-labelsets: An ensemble method for multilabel classification.* Paper presented at the European Conference on Machine Learning (ECML).

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory* (2nd ed.). New York: Springer-Verlag.

Wenemark, M., Borgstedt-Risberg, M., Garvin, P., Dahlin, S., Jusufbegovic, J., Gamme, C., . . . Björn, E. (2018). *Psykisk hälsa i sydöstra sjukvårdsregionen: En kartläggning av självskattad psykisk hälsa i Jönköping, Kalmar och Östergötlands län hösten 2015/16*. Retrieved from https://vardgivarwebb.regionostergotland.se/pages/285382/Psykisk_halsa_syostra_sjukvarsregionen.pdf

Zhang, M.-L., & Zhou, Z.-H. (2014). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering, 26*(8), 1819-1837.

Endnotes

1) The happy data are available upon request by contacting Marika Wenemark marika.wenemark@liu.se . The Immigrant and Civil Disobedience data are available from the Gesis Datorium http://dx.doi.org/10.7802/1795 . All data sets are available as of 1 August 2019.

2) The text was turned into n-gram variables in Stata using this statement for each data set

```
set locale_functions de  // German
ngram text , degree(1) binarize stemmer threshold(5) stopwords(.)
```

For the Swedish data the locale was set as follows:

```
set locale_functions sv   // Swedish
```

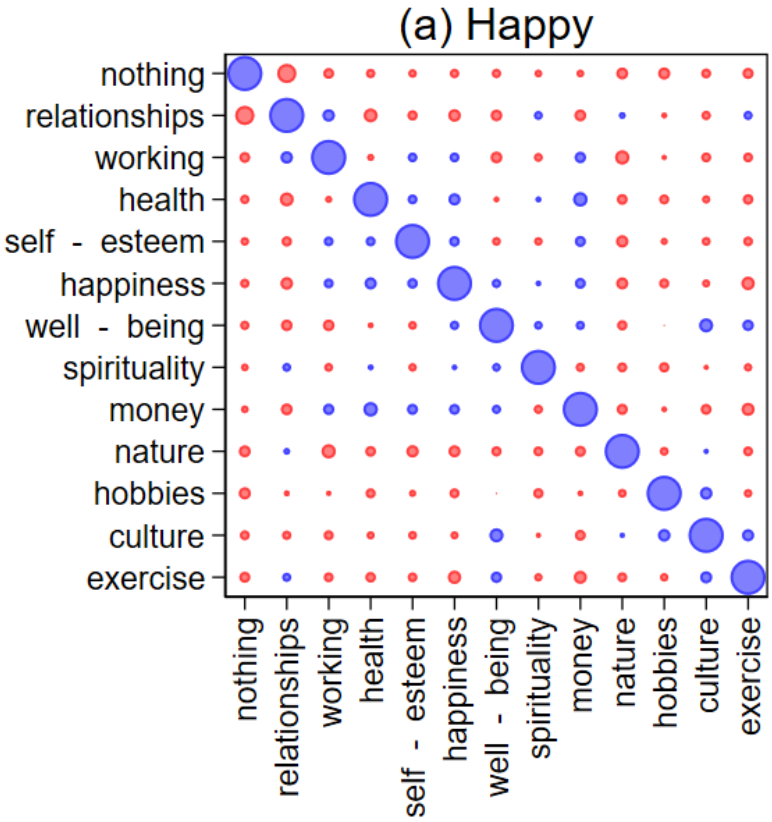
3) The R-code used to analyze the data is available in Appendix A in the Supplementary Material online.

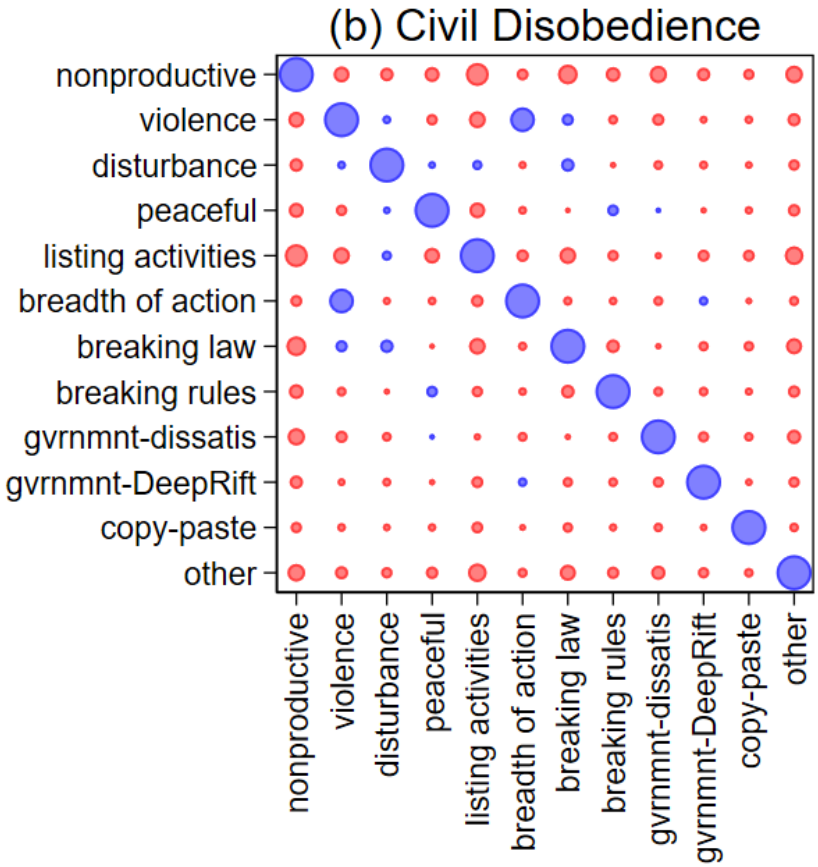
4) Figure 1 was produced in Stata with the program “corrplot.ado”, written by the first author. This can be found by typing “net search corrplot” in Stata.

	unigrams			bigrams		# words
text	x_emerg	x_go	x_room	x_emerg_room	x_go_emerg	
I am going to the emergency room	1	1	1	1	1	7

Table 1: Example of how a text is turned into n-gram variables (unigrams and bigrams)

Online version:




(b) Civil Disobedience
nonproductive
violence
disturbance
peaceful
listing activities
breadth of action
breaking law
breaking rules
gvrnmnt-dissatis
gvrnmnt-DeepRift
copy-paste
other

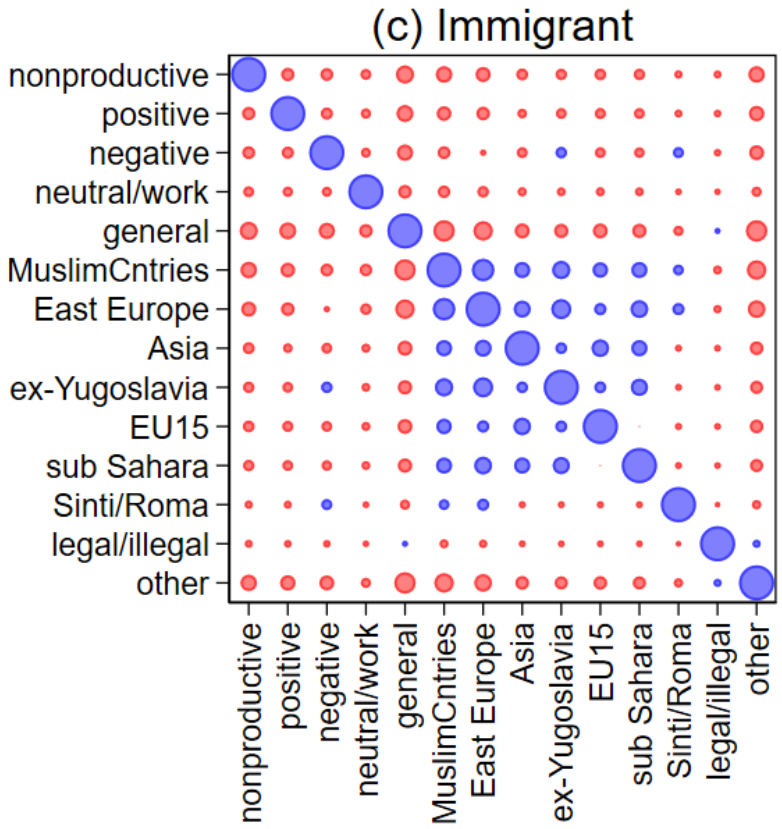
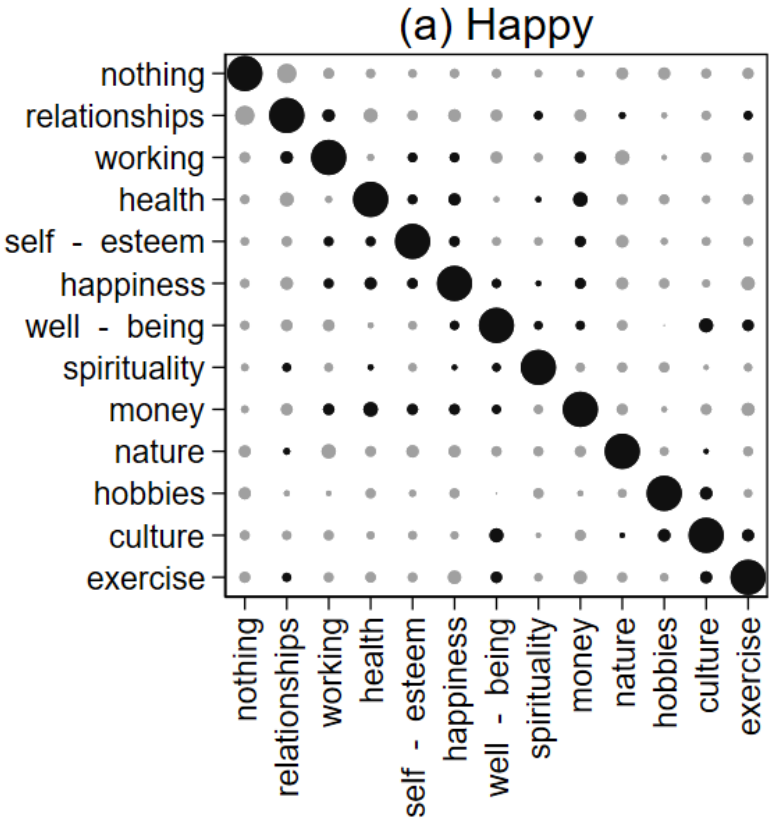
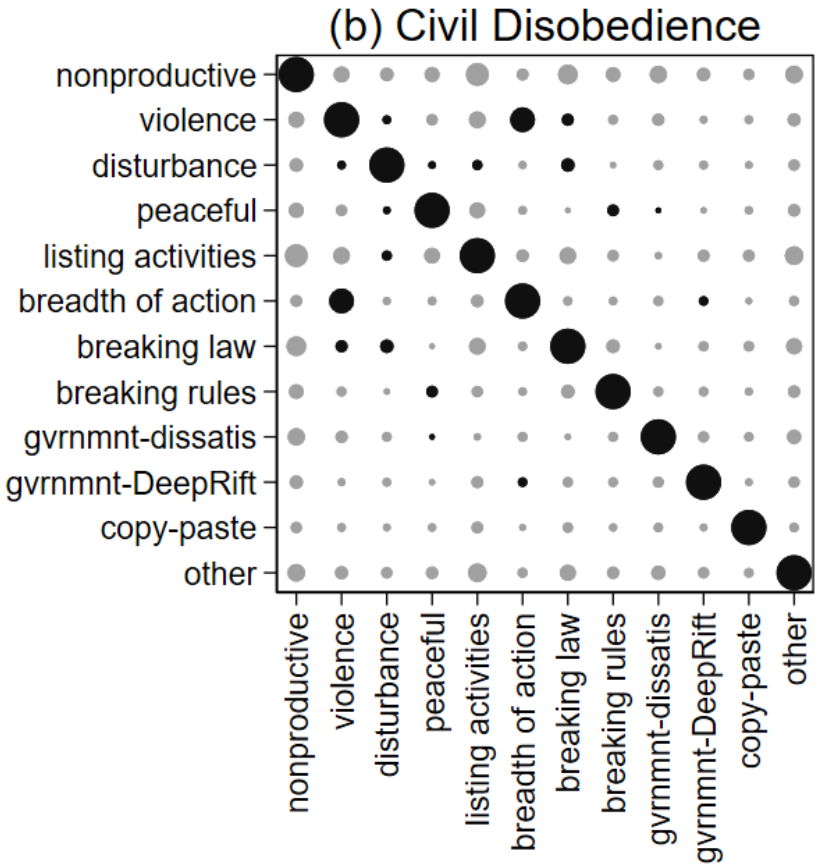


Figure 1: Plot of the correlation matrix of the three data sets. Each category on the x/y-axis corresponds to one of the labels.

Print version:




(b) Civil Disobedience
nonproductive
violence
disturbance
peaceful
listing activities
breadth of action
breaking law
breaking rules
gvrnmnt-dissatis
gvrnmnt-DeepRift
copy-paste
other

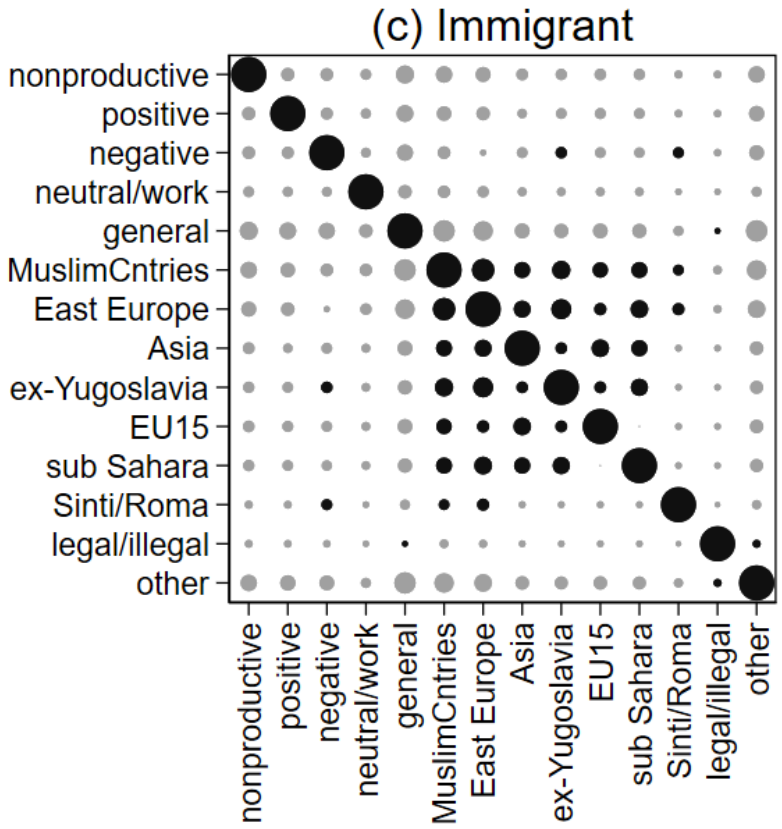


Figure 1: Plot of the correlation matrix of the three data sets. Each category on the x/y-axis corresponds to one of the labels. Blue (black) circles represent positive correlations, red (grey) circles represent negative correlations.

	av. # labels	# of labels	correlation				
			average	75th percentile	90th percentile	95th percentile	max
Happy	2.77	13	4.3%	5.6%	7.6%	10.5%	23.8%
Civil Disobedience	1.15	12	7.7%	9.4%	17.2%	22.4%	44.9%
Immigrant	1.44	14	8.6%	12.4%	19.2%	24.6%	35.7%

Table 2: Summary statistics of the bivariate label correlations by data sets.

	BR		ECC		RAKEL	
	0/1 loss	se	0/1 loss	se	0/1 loss	se
Happy	0.443	0.005	0.449	0.005	0.453	0.005
Civil Disobedience	0.524	0.009	0.465	0.009	0.478	0.009
Immigrant	0.474	0.008	0.358	0.007	0.385	0.007

Table 3: Percentage of respondents with one or more labels incorrectly classified (0/1 loss) by multi-label method. (Lower is better)

	BR		ECC		RAKEL	
	Hamming Loss	se	Hamming Loss	se	Hamming Loss	se
Happy	0.0506	0.0003	0.0538	0.0003	0.0532	0.0003
Civil Disobedience	0.0600	0.0006	0.0612	0.0006	0.0618	0.0006
Immigrant	0.0439	0.0004	0.0420	0.0003	0.0426	0.0003

Table 4: Percentage of labels incorrectly classified (Hamming loss) by multi-label method. (Lower is better).