Journal of Extension
www.joe.org

April 2018
Volume 56
Number 2
Article # 2TOT2
Tools of the Trade

Using Computational Text Classification for Qualitative Research and Evaluation in Extension

Abstract

This article introduces a process for computational text classification that can be used in a variety of qualitative research and evaluation settings. The process leverages supervised machine learning based on an implementation of a multinomial Bayesian classifier. Applied to a community of inquiry framework, the algorithm was used to identify evidence of cognitive presence, social presence, and teaching presence in the text contributions (44,000 unique posts) of more than 4,000 participants in an online environmental education course. Results indicate that computational text classification can significantly reduce labor costs and can help Extension research faculty scale, accelerate, and ensure reproducibility of their research.

Keywords: qualitative research, natural language processing, machine learning, text classification

Justin G. Smith
Assistant Professor
and County Extension
Director
justingriffis@wsu.edu

Reid Tissing
Software Specialist
reid.tissing@wsu.edu

Washington State
University
Shelton, Washington

Introduction

Processing open-ended interview and focus group responses and analyzing news reports and other texts are just some examples of qualitative research tasks undertaken in Extension (Smith & Lincoln, 1984). However, the unstructured character of the resulting data introduces significant complexity and cost in terms of data cleaning, processing, and analysis. Even with qualitative research software, analysis of medium to large volumes of text data can be time-consuming, highly subject to human error, and cost prohibitive. This situation poses constraints on the production and dissemination of Extension research and program evaluation to stakeholders and clients. Herein, we introduce a process Extension professionals can use to overcome such issues in a variety of qualitative research and evaluation settings. (See the appendix for definitions of terms used throughout the article.)

Computational text classification and other natural language processing (NLP) can facilitate large-scale as well as repeated qualitative analysis. Advances in computational linguistics and machine learning offer significant improvements in the efficiency and accuracy of NLP tasks, including text classification (Lai, Xu, Liu, & Zhao, 2015; Sebastiani, 2002), topic modeling (Řehůřek & Sojka, 2011), and sentiment analysis (Pang & Lee, 2008). Moreover, new algorithms and open source tools such as TextBlob (Loria, 2014), Natural Language Toolkit (NLTK), and Scikit-learn (Pedregosa et al., 2011) eliminate the need for the user to have knowledge of advanced mathematical concepts, enabling broader adoption and novel application of these tools.

Background and Data Collection

In early 2016, the Expanding Capacity in Environmental Education Project launched a massive open online course (MOOC) titled Environmental Education: Trans-disciplinary Approaches to Addressing Wicked Problems. Over 4,000 people participated in the 7-week professional development course. Participants interacted through the Canvas learning platform and a dedicated Facebook group. Between completing course assignments and engaging with others online, course participants generated over 44,000 unique posts.

The course evaluation team collected a random sample of posts by the MOOC participants and coded the sample using traditional qualitative coding methods (Saldaña, 2014) applied to a community of inquiry framework (Swan, Garrison, & Richardson, 2009). The evaluators coded responses on the basis of attributes related to cognitive presence, social presence, and teaching presence in the text of the MOOC posts. Because many of the responses were in essay format, coding time was significant. The coding task resulted in a set of 121 classified responses, with 91% interrater reliability.

Table 1 shows an example of the coding used to analyze course discussion posts for cognitive presence, which was indicated by four attributes: exploration, triggering events, integration, and resolution. To allow for better readability, we have included only three of the four indicators in Table 1, but the coding structure was the same for each indicator; that is, each indicator was given a value of 1 if a post contained evidence for the presence of the attribute or 0 if the post contained no evidence for the presence of the attribute.

Table 1.
Example of Course Discussion Posts Coded for Cognitive Presence

	Attributes indicative of cognitive presence		
Post	**Exploration**	**Triggering events**	**Integration**
What is our role in creating/enabling the continuation of/ preventing the resolution of wicked problems (through encouragement/ appeasement etc of certain policies)—environmentalists considered terrorists for trying to protect trees. . . .	0	1	0
Thank goodness for modern technology (despite another wicked problem with the issue of conflict minerals in the Congo http://www.enoughproject.org/conflict-minerals) to hold "law" enforcers accountable. Yet perhaps this event was needed to instil *[sic]* a sense of environmental and social justice for the young child in this video, who will be surely traumatized after witnessing the cognitive dissonance of watching people meant to	0	1	0

protect his community, hurting one of his community. Maybe he will grow up dedicated to unravelling wicked problems as a result? Who knows. Intriguing how this woman is put in cuffs and yet nobody from the mining companies responsible for the disaster in the Rio Doce, Brazil has been put in cuffs, or any of those responsible for triggering any number of environmental disasters.			
Yes!! The complicity of the State vs the rights of citizens vs the rights of trees!	1	0	0
"We" want cheap oil, high interest on our savings accounts etc. . . . how could the gov't regulate for that without cutting corners (pun intended)?	1	0	0

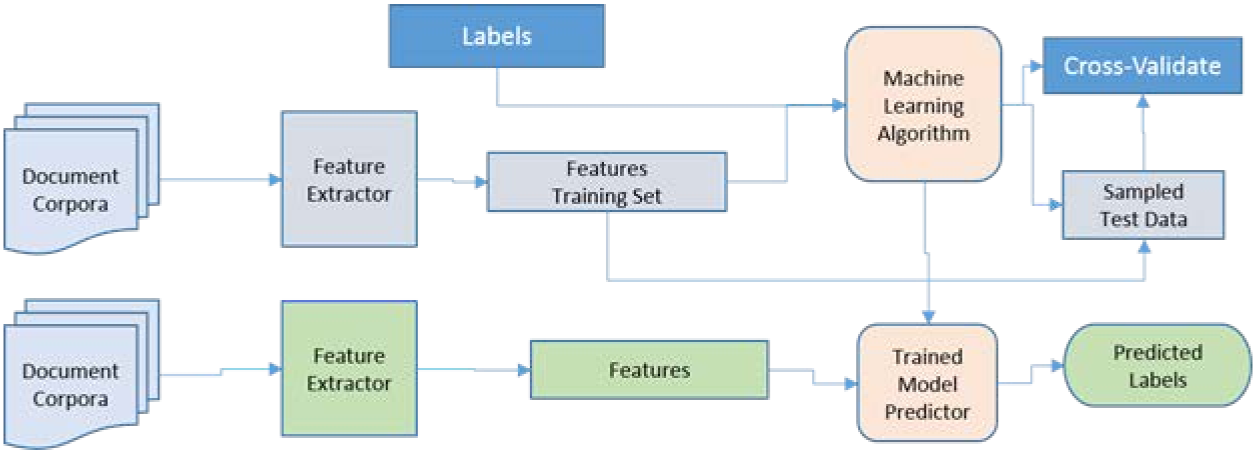
Although this approach is adequate for coding a small sample of data, analyzing data in larger amounts requires additional time and resources. Additionally, the thought of replicating the evaluation for future courses introduced further concerns regarding coding and reporting, in particular related to costs, efficiency, and accuracy.

Enter Text Classification

As a result of the aforementioned concerns, our team was brought in to support a larger analysis that could be both scalable and reusable. With our approach, we combined computational text classification and machine learning to predict labels (categories) associated with each text entry. We used a supervised classification method and previously coded responses as inputs to "train" a predictive model. Figure 1 depicts a common workflow for supervised text classification, where each document (or text entry) is partitioned into a set of features and those features are associated with the user-defined categories (labels). The "training" data are used to produce a probability distribution that associates a given feature set with a set of labels; this becomes our "model" and is used to fit new (unlabeled) data according to the feature-label probability distribution.

Figure 1.

Process Model for Supervised Text Classification



Implementing the Classifier

With the MOOC participant data, we used the baseline Bayes method provided out-of-the-box from Scikit-learn and TextBlob. Ease of implementation and robust results make the Bayesian method the typical starting point in any machine learning system, and it is often used as the benchmark for evaluating other algorithms. Bayes's theorem is characterized as "naive" because it is built on the assumption of independence between every pair of possible features. Thus, given the variable *y* and a dependent feature vector x_1 through x_n, Bayes's theorem states the following relationship:

$$P(y \mid x_1,\ldots,x_n) = \frac{P(y)P(x_1,\ldots x_n \mid y)}{P(x_1,\ldots,x_n)}$$

Although math is involved, tools such as TextBlob provide simple-to-use interfaces for working with both NLTK text tools and Scikit-learn predictive models. Simple Python import statements (e.g., "import textblob") allow the user to implement the NLTK and Scikit-learn methods, and new methods can be implemented on top of these base packages.

We organized texts, formatted them into CSV files, and read them into memory one document at a time. We processed each document to calculate the frequency distribution for the prelabeled data on the basis of common features across the collection of documents. The training time for feature detection was about 5 min on a training set consisting of 121 documents. Once the classifier was trained, uncoded data (test data) were evaluated against the trained model; labels were generated for each document on the basis of probability scores of relatedness between the preestablished labels and features found in the document. Label results were evaluated using NLTK's built-in accuracy method and were cross-checked by the evaluation team. We found that the process resulted in accurate label predictions for 91% of the test data.

Our results indicate that the approach described herein can be used to reduce the time needed to conduct deep analysis of qualitative data, but they also show that human oversight continues to be necessary. This is particularly the case when researchers are attempting to fit a data set into a predefined theoretical construct. Other algorithms could be applied as well; k-means clustering highlighted in Skelly, Hill, and Singletary (2014), support vector machines, and neural nets have all shown success in text classification contexts. Furthermore, such approaches could be used to extend ideas presented by Skelly et al. (2014) for massive-scale analysis as shown by Diao, Erkan, Hassan, and Radev (2011).

Conclusion

NLP and machine learning are rapidly becoming important approaches to managing and using the vast volumes of data being created across the web. Although NLP does not eliminate the need for people to evaluate data, it does offer Extension professionals and researchers opportunities to accelerate and perform data analysis on a larger scale. It also offers a path for using previously coded qualitative data as training data across a range of unstructured data sets. And with new tools coming online, such as Text Aylien (http://aylien.com/) and others, researchers can begin to leverage language processing models from within Google Sheets or custom interfaces that eliminate the need for programming altogether. Overall, the combination of human and machine learning approaches to data classification can be used to strengthen and broaden the scope of qualitative research output across Extension.

References

Diao, Q., Erkan, G., Hassan, A., & Radev, D. R. (2011). *Improved nearest neighbor methods for text classification.* Retrieved from https://www.semanticscholar.org/paper/Improved-Nearest-Neighbor-Methods-For-Text-Erkan-Hassan/f6783875b62d1eecbb0160d1cc12e26b50e81612

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. *Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, *333*, 2267–2273.

Loria, S. (2014). *TextBlob: Simplified text processing*. Retrieved from https://textblob.readthedocs.io/en/dev/

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, *2*(1–2), 1–135.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Řehůřek, R., & Sojka, P. (2011). *Gensim–Python framework for vector space modelling*. Brno, Czech Republic: Natural Language Processing Centre, Faculty of Informatics, Masaryk University.

Saldaña, J. (2014). *The coding manual for qualitative researchers*. London, UK: Sage Publications.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys* (CSUR), *34*(1), 1–47.

Skelly, J., Hill, G., & Singletary, L. (2014). Probing needs assessment data in depth to target programs more effectively. *Journal of Extension*, *52*(2), Article 2RIB1. Available at: https://www.joe.org/joe/2014april/rb1.php

Smith, M. F., & Lincoln, Y. S. (1984). Another kind of evaluation. *Journal of Extension*, *22*, Article 6FEA1. Available at: https://www.joe.org/joe/1984november/a1.php

Swan, K., Garrison, D. R., & Richardson, J. (2009). A constructivist approach to online learning: The community of inquiry framework. In C. R. Payne (Ed.), *Information technology and constructivism in higher education: Progressive learning frameworks* (pp. 43–57). Hershey, PA: IGI Global.

Appendix

Glossary of Terms

Computational Linguistics: The branch of linguistics in which the techniques of computer science are applied to the analysis and synthesis of language and speech.

Computational Text Classification: Digital articulation of meaning embedded within a text or corpus.

Data Reduction: The transformation of numerical or alphabetical digital information derived empirically or experimentally into a corrected, ordered, and simplified form.

Feature Vector: An n-dimensional vector of numerical features that represent some object, which facilitates processing and statistical analysis.

Labels: Human-created classifiers containing a meaningful "tag" (informational feature) or the resulting output of classifiers generated by machine learning analysis.

Label Prediction: The accuracy of premade labels' occurrence contained in the generated output of an algorithmic analysis.

Machine Learning: The ability of computer software to analyze data by finding patterns and using this information to learn without human intervention. Usually categorized as supervised, semisupervised, or unsupervised.

Natural Language Processing/NLP: A method for computers to analyze, understand, and derive meaning from human language to perform tasks such as automatic summarization relationship extraction, sentiment analysis, speech recognition, and topic segmentation.

Neural Networks: Computing systems based on the biological framework of the human brain, which use layers made up of interconnected nodes to "learn" by processing connections between nodes to generate an output based on patterns found from the initial input.

Sentiment Analysis: The use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.

Support Vector Machines/SVM: Supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other.

Test Data: A set of data used to assess the strength and utility of a predictive relationship.

Text Corpus: A large and structured set of texts, used for conducting statistical analysis, testing hypotheses, checking occurrences, or validating linguistic rules within a specific language territory.

Topic Model: A type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body.