Text as Data

Kohei Watanabe

In-press (submitted in April 2021)

For *Encyclopedia of Technology & Politics* by Edward Elgar Publishing.

Introduction

The term "text as data" (Grimmer & Stewart, 2013) refers to quantitative analysis of digitalized texts to reveal properties of authors or the content of documents such as speech transcripts, election advertisements, government reports, newspaper articles, and social media posts. In political research, the methodology has been employed to measure, for example, the ideology of election candidates (Catalinac, 2018), the fairness of treaties (Spirling 2012), the originality of bills (Jansa et al., 2018), the bias in news reporting (Watanabe, 2017), the diversity of news content (Amsalem et al., 2020) and the intensity of geo-political threat (Trubowitz & Watanabe, 2021).

Texts are sequence of words, but researchers can analyzed them statistically with natural language processing techniques. These techniques are implemented and made accessible through various software packages and libraries. The shift from print to online publishing and digitalization of documents in recent years made it possible for researchers to apply the methodology to a wide range of political documents. The increased capacity of personal computers to process large datasets also made quantitative text analysis more accessible. The latest development in machine learning techniques is making quantitative text analysis more useful, but it is also raising concerns and posing new challenges for researchers such as the disconnection between analysis and theory and the high cost to sufficiently train complex models.

Historical overview

The root of quantitative text analysis can be traced back to the 1960s, when The General Inquirer (Stone et al., 1966) was created for dictionary analysis. The program searches the document for words that are registered in a pre-defined list of keywords called "dictionary". In the following years, various dictionaries such as the Psychosociological Dictionary, Political Dictionary, and Anthropological Dictionary were created to answer specific theoretical questions. The General Inquirer, which runs on mainframe computers, was not widely used but it prompted development of similar software programs for

dictionary analysis, which run on personal computers.[1] Most importantly, the General Inquirer dictionaries are combined and modified to construct dictionaries such as the Linguistic Inquiry and Word Count (LIWC) and the Lexicoder Sentiment Dictionaries, which are still used by political scientists today.

Quantitative text analysis in political science attracted broader attention in the 2000s, when new statistical analysis methods and data processing tools became available. Wordscores (Benoit & Laver, 2003) and Wordfish (Slapin & Proksch, 2008) were proposed to estimate ideological positions of authors; the Latent Dirichlet Allocation (Blei et al., 2003) and the Structural Topic Model (Roberts et al., 2014) were created to identify salient topics in documents. Open-source software packages for textual data processing and analysis were also developed in R or Python to apply these methods.[2]

The development of the methodology reflects the recent advancement in natural language processing technologies and machine learning. The Unicode covers most of the languages and its library processes multi-lingual documents reliably and consistently; machine learning models can replicate human judgement accurately; neural networks, especially the deep learning models, further enhance the human-like ability of machine learning models.[3] Conventional machine learning models have been more common in quantitative text anlaysis, but neural networks have also been employed in recent works.[4]

Characteristics of textual data

Researchers employ a distinctive set of tools in quantitative text analysis because textual data in its original form is an unstructured sequence of symbols. This unique characteristic of textual data requires conversion of symbols into numbers and reduction of noises before applying statistical analysis methods. The preprocessing of texts results in two different forms, the "bag-of-words" or "string-of-words" representation, depending on the level of simplification.

In preparation for statistical analysis, textual data is aggressively simplified based on the symbols or frequencies as feature selection. Symbol-based simplification includes deletion of common grammatical

[1] Software programs for dictionary-based analysis includes LIWC (Francis & Pennebaker, 1993), DICTION (North et al., 1999), Yoshicoder (Lowe, 2006) and Lexicoder (Young & Soroka, 2012).
[2] Software packages for quantitative text analysis includes tm (Feinerer et al., 2008), tidytext (Silge & Robinson, 2016), quanteda (Benoit et al., 2018), NLTK (Bird et al., 2009), Gensim (Rehurek & Sojka, 2010).
[3] A greater number of neural network models are available in Python libraries that include Keras, Scikit Lean and PyTorch.
[4] See Wankmüller (2021) for an overview of neural network models in social research. Recent application in political research can be found in Zhang & Pan (2019), Chang & Masterson (2020) and Muchlinski et al. (2021).

words (“stop words”), mechanical truncation of word endings (“stemming”), and substitution by uninflected forms of words (“lemmatization”); frequency-based simplification includes removal of very frequent or infrequent words. The dictionary analysis is also a type of symbol-based simplification of textual data.

Bag-of-words representation

Texts can be represented in a matrix (“document-feature” matrix) whose rows correspond to documents and whose columns correspond to unique words (“types”). This form of textual data is called “bag-of-words” because the positions of words are discarded entirely. However, the bag-of-words representation is often sufficient for statistical analysis because it still records frequency of words in each document, the entire collection (“corpus”), and their co-occurrences in documents.

In the bag-of-words representation, frequent or infrequent word can be easily identified by their overall frequencies (“term frequency”) or the number of documents the words occur (“document frequency”). It is also common to apply statistical measure such as Shannon entropy to select informative features.

String-of-words representation

Texts can also be represented as vectors of numeric identifiers (“tokens”) which corresponds to words in documents. This form of textual data can be called “string-of-words” because the relative positions of original words are preserved. With the sequence of words, statistical models can consider semantic or syntactical relationship between them to perform tasks more accurately. Many of the recently developed neural network models requires this form of data representation.

The string-of-words representation requires larger storage space and greater computing power than the bag-of-words representation. Further, the simplification of data is also limited to symbol-based approaches as it does not offer information on frequency of words. Therefore, the size of documents (or the window) for statistical analysis must be small to avoid exponentially increase in computational costs.

Analysis of textual data

Researchers can find language phenomena in different levels using the methodology. The goals of analysis are, for example, revealing topic or sentiment of documents on the *discourse level*, identifying synonyms or antonyms on the *semantic level*, and extract multi-word expressions such as phrases on the *lexical level*.[5] The discourse-level analysis is often the most important in political research, but semantic or lexical-level analysis is also performed as part of preprocessing data. Semantic-level analysis is

[5] See Liddy (2001) for more detailed discussion on language phenomena in natural language processing.

performed using word-embedding models that compute semantic similarity accurately; lexical-level analysis is achieved by collocation analysis that discover strongly associated sequences of words.[6]

Depending on the goals of the analysis and the available resources, researchers choose different approaches to textual data. Dictionary analysis is a symbolic and knowledge-based approach; both statistical analysis and machine learning are numeric and data-driven approaches, but the main goals are describing the relationship between words and documents in the statistical analysis while predicting classes or positions of documents in the machine learning.

The results of the analysis are scrutinized by researchers to ensure the validity of the measurement. They should have patterns that consistent not with the existing knowledge ("face valid") and values that correlates with the gold-standard ("criterion valid"), which are often created based on manual analysis.

Dictionary analysis

Dictionary analysis is still very popular for its technological simplicity. A dictionary is a long list of words that are nested under categories relates to concepts of interests. Researchers receive frequencies of categories (or keys) instead of original words from dictionary analysis. The result can be either manually interpreted or passed to statistical analysis methods.

Researchers can produce meaningful results easily with a dictionary in both large and small collection of documents, because it is solely based on the pre-defined list of words. Several dictionaries are publicly available but use of them requires caution because validity of dictionary analysis depends on their suitability in target domains (Boukes et al., 2020; Grimmer & Stewart, 2013).[7] When suitable dictionaries are unavailable, researchers must compile their own dictionary based on their knowledge of the content.[8]

Statistical analysis

After appropriate preparation of data, statistical analysis is performed to find association between words and documents. Simple frequency analysis or relative frequency analysis can be applied to documents to

[6] Popular word-embedding models are Latent Semantic Analysis (Deerwester et al., 1990) and Global Vector (Pennington et al., 2014). Collocation analysis can be performed by two-by-two association measure such as point-wise mutual information (PMI), chi-squared or log-likelihood statistic (Hoey, 2012).
[7] The English language has the largest collection of dictionaries that include the Regressive Imagery Dictionary (Martindale, 1975), the Lexicoder Sentiment Dictionary (Young & Soroka, 2012), the Moral Foundation Dictionary (Frimer et al., 2017), and Affective Norms for English Words (Nielsen, 2011).
[8] Published dictionaries usually contain hundreds of words, but corpus specific dictionaries can produce valid results with small number of words (Müller, 2020). See Watanabe and Zhou (2020) on the best practice in creating original dictionaries.

understand the discourse.[9] Collocation analysis identifies strongly associated sequences of multi-word expressions that includes phrases or proper names; identified multi-word expressions can be used to compound words to improve the bag-of-words representation of the data before applying other statistical analysis or machine learning methods.

Clustering techniques can also be applied to group documents or words based on their discursive or semantic similarities. For example, high similarity between documents indicate that they are about the same subject or by the same author (Jansa et al., 2018); changes in similarities between words reveals shifts in discourse (Rodman, 2020). Similar documents can also comprise a composite document in further statistical analysis; similar words can enter a dictionary as synonyms (Watanabe, 2020).[10]

Machine learning

Machine learning techniques are employed to automatically assign documents into classes (e.g. topics) or position them on a continuous scale (e.g. sentiment). Supervised learning algorithms replicate labeling or scoring of documents based on training on human-coded instances. Unsupervised learning algorithms distinguish between documents based on distribution of words without training instances. Semi-supervised learning algorithms distinguish between documents based on both distribution of words and human-provided keywords.[11]

The result of document classification or scaling by machine learning techniques depends not only on the algorithm and data preparation, but also the quality and the size of human inputs. Supervised learning models must be trained on a sufficiently large number instances to perform the tasks accurately. Unsupervised learning models requires careful selection of features and hyper-parameters to produce meaningful results. Semi-supervised learning models demands unambiguous operationalization of concepts by keywords in a similar way as dictionary analysis.

---

[9] Signed-chi squared statistic (or Keyness) computed by based on relative frequencies of words in two groups of documents helps researchers to find meaningful words (Bondi, 2010).

[10] Document similarity is usually computed as cosine or Euclidian similarity between a pair of vectors that records frequencies of words ("document vectors"), while word similarity is similarity between a pair of vectors that records their frequencies in documents ("word vectors").

[11] For example, Naïve Bayes classifier, Wordscores (Benoit & Laver, 2003), support-vector machines and random trees are supervised learning models; Wordfish (Slapin & Proksch, 2008), Latent Dirichlet Allocation (Blei et al., 2003) and Structural Topic Model (Roberts et al., 2014) are unsupervised learning models; Seeded-LDA (Lu et al., 2011), Keyword-Assisted Topic Model (Eshima et al., 2020) and Latent Semantic Scaling (Watanabe, 2020) are semi-supervised learning models.

Conclusions

Quantitative text analysis has developed rapidly in the last 20 years thanks to the greater availability of software packages and libraries for natural language processing and new analysis techniques. However, there are issues that need to be addressed to further develop political research with the methodology.

The main issues are the understudying of documents in non-English languages and the disconnection between machine learning results and political science concepts (Eshima et al., 2020; Watanabe & Zhou, 2020). Political scientists can process non-English language texts with the recently developed software tools but cannot analyze them without lexical resources such as stopwords lists and dictionaries. Machine learning techniques has the great potential, but the state-of-art supervised learning models, especially deep learning models, often require too much manual inputs to train; unsupervised machine learning models do not always produce results meaningful to researchers.

The lack of non-English lexical resources can be partially addressed by translating an English-language lexicon list into other languages.[12] The connection between machine learning results and political science concepts can be reestablished by the semisupervided learning models that are considerably less expensive than fully supervised learning models; recent develop techniques to repurpose trained models ("transfer learning") would also reduce the cost. Once these issues are address, researchers will be able to perform cross-national studies using quantitative text analysis and further enrich our knowledge of politics.

References

Amsalem, E., Fogel-Dror, Y., Shenhav, S. R., & Sheafer, T. (2020). Fine-Grained Analysis of Diversity Levels in the News. *Communication Methods and Measures*, *14*(4), 266–284. https://doi.org/10.1080/19312458.2020.1825659

Benoit, K., & Laver, M. (2003). Estimating Irish party policy positions using computer wordscoring: The 2002 election – a research note. *Irish Political Studies*, *18*(1), 97–107. https://doi.org/10.1080/07907180312331293249

[12] Proksch et al. (2019) translated the Lexicoder Sentiment Dictionary into 20 European languages; Matsuo et al. (2019) created a Japanese-language version of the Moral Foundation Dictionary. Marimo stopwords lists (https://github.com/koheiw/marimo) that covers Asian languages such as Arabic, Hebrew, Korean, and Chinese.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022. https://doi.org/10.5555/944919.944937

Bondi, M. (2010). Perspective on keywords and keyness: An introduction. In M. Bondi & M. Scott (Eds.), *Keyness in texts* (pp. 1–18). John Benjamins.

Boukes, M., van de Velde, B., Araujo, T., & Vliegenthart, R. (2020). What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools. *Communication Methods & Measures*, *14*(2), 83–104. https://doi.org/10.1080/19312458.2019.1671966

Catalinac, A. (2018). Positioning under Alternative Electoral Systems: Evidence from Japanese Candidate Election Manifestos. *American Political Science Review*, *112*(1), 31–48. https://doi.org/10.1017/S0003055417000399

Chang, C., & Masterson, M. (2020). Using Word Order in Political Text Classification with Long Short-term Memory Models. *Political Analysis*, *28*(3), 395–411. https://doi.org/10.1017/pan.2019.46

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, *41*(6), 391–407.

Eshima, S., Imai, K., & Sasaki, T. (2020). Keyword Assisted Topic Models. *ArXiv:2004.05964 [Cs, Stat]*. http://arxiv.org/abs/2004.05964

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, *25*(5), 1–54.

Francis, M. E., & Pennebaker, J. W. (1993). *LIWC: Linguistic Inquiry and Word Count* [Technical Report]. Southern Methodist University.

Frimer, J., Haidt, J., Graham, J., & Dehgani, M. (2017). *Moral Foundations Dictionaries for Linguistic Analyses, 2.0*. http://www.jeremyfrimer.com/uploads/2/1/2/7/21278832

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Hoey, J. (2012). *The Two-Way Likelihood Ratio (G) Test and Comparison to Two-Way Chi Squared Test*. http://arxiv.org/abs/1206.4881

Jansa, J. M., Hansen, E. R., & Gray, V. H. (2018). Copy and Paste Lawmaking: Legislative Professionalism and Policy Reinvention in the States. *American Politics Research*, 1532673X18776628. https://doi.org/10.1177/1532673X18776628

Liddy, E. D. (2001). Natural language processing. In M. Dekker (Ed.), *Encyclopedia of Library and Information Science* (2nd edition). http://surface.syr.edu/cgi/viewcontent.cgi?article=1019&context=cnlp

Lowe, W. (2006). Yoshikoder: An open source multilingual content analysis tool for social scientists. *Annual Meeting of the American Political Science Association, Philadelphia, PA*.

Lu, B., Ott, M., Cardie, C., & Tsou, B. K. (2011). Multi-aspect sentiment analysis with topic models. *2011 IEEE 11th International Conference on Data Mining Workshops*, 81–88.

Martindale, C. (1975). *Romantic progression: The psychology of literary history*. Hemisphere Publishing.

Matsuo, A., Sasahara, K., Taguchi, Y., & Karasawa, M. (2019). Development and validation of the Japanese Moral Foundations Dictionary. *PLOS ONE*, *14*(3), e0213343. https://doi.org/10.1371/journal.pone.0213343

Muchlinski, D., Yang, X., Birch, S., Macdonald, C., & Ounis, I. (2021). We need to go deeper: Measuring electoral violence using convolutional neural networks and social media. *Political Science Research and Methods*, *9*(1), 122–139. https://doi.org/10.1017/psrm.2020.32

Müller, S. (2020). Media Coverage of Campaign Promises Throughout the Electoral Cycle. *Political Communication*, *37*(5), 696–718. https://doi.org/10.1080/10584609.2020.1744779

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *ArXiv:1103.2903 [Cs]*. http://arxiv.org/abs/1103.2903

North, R., Lagerstrom, R., & Mitchell, W. (1999). *DICTION Computer Program* (Inter-University Consortium for Political and Social Research). University of Michigan. http://www.icpsr.umich.edu.gate2.library.lse.ac.uk/icpsrweb/ICPSR/studies/5909/version/1

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Proksch, S.-O., Lowe, W., Wäckerle, J., & Soroka, S. (2019). Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches. *Legislative Studies Quarterly*, *44*(1), 97–131. https://doi.org/10.1111/lsq.12218

Rehurek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Roberts, M., Stewart, B., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S., Albertson, B., & Rand, D. (2014). Structural topic models for open ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082.

Rodman, E. (2020). A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors. *Political Analysis*, *28*(1), 87–111. https://doi.org/10.1017/pan.2019.23

Silge, J., & Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *JOSS*, *1*(3). https://doi.org/10.21105/joss.00037

Slapin, J. B., & Proksch, S.-O. (2008). A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, *52*(3), 705–722. https://doi.org/10.1111/j.1540-5907.2008.00338.x

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Trubowitz, P., & Watanabe, K. (2021). The Geopolitical Threat Index: A Text-Based Computational Approach to Identifying Foreign Threats. *International Studies Quarterly*, *sqab029*. https://doi.org/10.1093/isq/sqab029

Wankmüller, S. (2021). Neural Transfer Learning with Transformers for Social Science Text Analysis. *ArXiv:2102.02111 [Cs, Stat]*. http://arxiv.org/abs/2102.02111

Watanabe, K. (2017). Measuring news bias: Russia's official news agency ITAR-TASS' coverage of the Ukraine crisis. *European Journal of Communication*, *32*(3), 224–241. https://doi.org/10.1177/0267323117695735

Watanabe, K. (2020). Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages. *Communication Methods and Measures*. https://doi.org/10.1080/19312458.2020.1832976

Watanabe, K., & Zhou, Y. (2020). Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches: *Social Science Computer Review*. https://doi.org/10.1177/0894439320907027

Young, L., & Soroka, S. (2012). Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication*, *29*(2), 205–231. https://doi.org/10.1080/10584609.2012.671234

Zhang, H., & Pan, J. (2019). CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media. *Sociological Methodology*, *49*(1), 1–57. https://doi.org/10.1177/0081175019860244