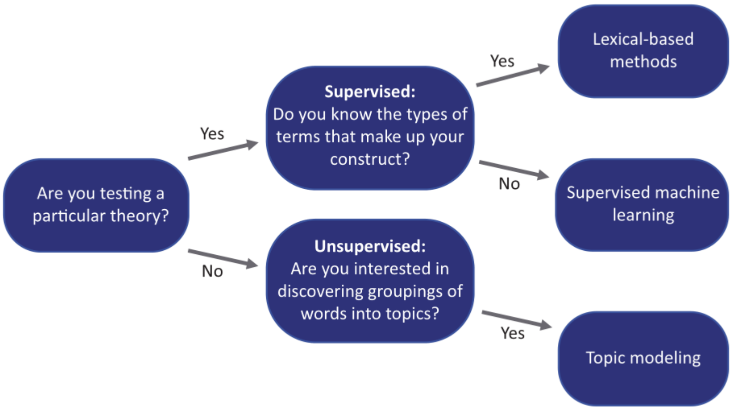
Text analysis method in education

1 Overview


Are you testing a particular theory?
Yes
No
Supervised:
Do you know the types of terms that make up your construct?
Unsupervised:
Are you interested in discovering groupings of words into topics?
Yes
No
Yes
Lexical-based methods
Supervised machine learning
Topic modeling

1.1 Situation

1. More than ever in the past, human have access to broad, rich, educationally–relevant text data from sources such as online discussion forums, transcribed video and audio of face–to–face classes, digital essays, social media, and emails and texts between parents and schools.
2. Computational linguists and social scientists have recently developed more advanced and nuanced tools that can be applied to quantitatively analyze large–scale text data.

1.2 Application

1. The advances in data availability and analytic techniques can dramatically expand our capacity for discovering new patterns and testing new theories in education.
2. They can also expand the types of research questions that researchers can ask, improve the external validity and representativeness of research, and reduce the cost to complete these types of studies.

2 Lexical–Based Methods

2.1 Situation

- If researchers know the types of terms that make up their construct, lexical–based methods work best.
- These methods rely on the number of times words occur in documents to measure the construct of interest.

2.2 Tools

2.2.1 Lists of Words

In some researchs, the researchers were able to determine their lists of words from their own data. However, despite knowing the types of terms that make up their construct, researchers may not always be able to easily construct a word list from their own data.

2.2.2 Linguistic Inquiry and Word Count (LIWC)

- LIWC is the most widely used dictionary-based sentiment analysis tool used among social scientists, and its word lists were constructed through a combination of using existing dictionaries,thesauruses, common emotion rating scales, and human-generated lists, and ratings by judges.
- Each of the words on the LIWC lists are weighted equally.

2.2.3 Sentiment Analysis and Social Cognition Engine (SEANCE)

- SEANCE is a newer tool that combines lists of words from eight open source databases with tools that identify parts of speech and negations. This allows SEANCE to analyze adjectives separately from nouns (for example), and to ignore terms with negations (e.g. not) before them. SEANCE also includes summary indices produced from a principal componentanalysis on a corpus of movie reviews from the Internet Movie Database (IMDb).
- Each of the words from SEANCE are thus weighted based on their principal component loadings.

2.2.4 Compare the dictionary methods to hand coding

- Despite the ease of use of external dictionaries (and their correspondingly wide application), they can perform poorly when applied to a different domain.
- The best way to validate dictionaries is to compare the dictionary methods to hand coding.

2.4 Cases

Baker, Bloom, and Davis (2016) .Measuring economic policy uncertainty.
They determine whether newspaper articles discuss policy uncertainty based on whether the articles contain terms like “regulation,” “deficit,” and “federal reserve;”

Evans, Marsicano, and Lennartz (2019) Cracks in the Bedrock of American Democracy: Differences in civic engagement across institutions of higher education.
They assess postsecondary institutions’ commitment to civic engagement by observing whether terms such as “volunteer” and “service” exist in the institutions’ mission statements;

Bettinger, Liu, and Loeb (2016).Connections matter: How interactive peers affect students in online college courses.

They measure peer interactivity in online classes by constructing a course-specific roster and determining whether each post contained a name on that course's roster or not.

Quinn et al. (2010) .How to analyze political attention with minimal assumptions and costs
They correlate their measure of Congress' discussion of abortion to official abortion roll-call votes.

Baker et al. (2016) . Measuring Economic Policy Uncertainty.
Correlate their measure of economic policy uncertainty to an established measure of volatility in the S&P500 index.

3 Supervised Machine Learning

3.1 Situation

Researchers know what theory they would like to test but do not know which words or linguistic features make up their construct.

3.2 Steps

1. Hand-code a subset of the total documents into the categories of interest
2. A document-term matrix (evaluate the reliability of the hand-coded labels)
3. 【Build their model using one subset of the data (the training data)】Use the relationships between document-level features and their hand-coded categories to predict the categories for unlabeled documents.
4. Assess model performance on another subset of the data【k-fold cross validation】

3.3 Main methods

Researchers should consider using multiple types of machine learning algorithms to assess which maximizes performance.

1. ordinary least squares
2. logistic regression
3. regularized regression like LASSO, ridge, or elastic net
4. Support Vector Machine
5. decision tree algorithms like random forests
6. a deep learning method like a neural net

3.4 Cases

Using machine learning to translate applicant work history into predictors of performance & turnover
This study used supervised machine learning to classify teacher applicants' self-reported reasons for leaving their previous job into four categories (involuntary, avoiding bad jobs, approaching better jobs, and other reasons). They hand-coded 1,000 of the self-reported reasons, then were able to train a model using those 1,000 observations to predict the reasons for the remaining 35,000 applicants. They could then use these

predictions to study the relationship between teacher turnover reasons and job performance.

Automatically measuring question authenticity in real-world classrooms
This study used supervised machine learning to predict when teachers ask questions without predetermined answers by building a model based on a set of hand-coded questions.

Does feedback matter? Performance management and improvement in public education
Researchers hand-code teacher observation feedback into seven major domains for a subset of teachers, creating a foundation for using supervised machine learning to predict the domains for all teachers in Tennessee.

4 Unsupervised Machine Learning

4.1 Situation

Researchers would like to discover new theories or topics in their data
In social science, mainly ”topic models”

4.2 Application

1. Topic models use the terms in the documents (which are observable) to determine the topics being discussed (which are latent and unobservable)——The objective of the topic model is to find the parameters that have generated bag-of-words documents.
2. Topic modeling can also be used to explore how content varies by document characteristics. [STM]

4.3 Steps

1. Estimate the topic model
2. Researchers examines the topic groupings andmanually labels the topics

3.2 Cases

Differing views of equity: How prospective teachers perceive their role in closing achievement gaps
This study used topic models on essay responses of 10,000 teachers to examine how teachers perceive their role in closing achievement gaps. Topic modeling allowed them to examine essay content without imposing their own groupings on the data and led to the discovery that that Hispanic and African American applicants discuss structural causes of inequality more frequently.

The civic mission of MOOCs: Measuring engagement across political differences in forums
This study used topic models to explore liberal and conservative students’ language in MOOCs and find that these groups of students use largely the same language to discuss similar topics (like the Common Core and school vouchers).

Computer-assisted reading and discovery for student-generated text in massive open online courses

This study used an STM to estimate how student motivations in online courses vary by gender and found that male students were more likely to state that they enrolled in a class because of the university’s elite association than female students were.