Text Analysis: An Introduction with R

Social Science Data Analytics Workshop

Caleb Lucas (@calebjlucas)

July 23, 2020

Michigan State University

Introduction

- A sentence (or two) about me

Introduction

- A sentence (or two) about me
- Goals for today:

Introduction

- A sentence (or two) about me
- Goals for today:
  - Brief introduction to several text analysis methods

Introduction

- A sentence (or two) about me
- Goals for today:
    - Brief introduction to several text analysis methods
    - Broad overview

Introduction

- A sentence (or two) about me
- Goals for today:
    - Brief introduction to several text analysis methods
    - Broad overview
        - A restaurant menu, not a cooking class

Introduction

- A sentence (or two) about me
- Goals for today:
    - Brief introduction to several text analysis methods
    - Broad overview
        - A restaurant menu, not a cooking class
        - Very little/no math

Text Analysis- what is it?

Text Analysis

- Text analysis uses computers and statistics to extract information (patterns, entities, topics, etc) from text

Text Analysis

- Text analysis uses computers and statistics to extract information (patterns, entities, topics, etc) from text
- We typically use text to make inferences about some latent variable

Text Analysis

- Text analysis uses computers and statistics to extract information (patterns, entities, topics, etc) from text
- We typically use text to make inferences about some latent variable
    - Observe transcripts, news articles, social media posts, etc.

Text Analysis

- Text analysis uses computers and statistics to extract information (patterns, entities, topics, etc) from text
- We typically use text to make inferences about some latent variable
    - Observe transcripts, news articles, social media posts, etc.
    - Make inferences about things we can't directly observe like ideology

Text Analysis

- Text analysis uses computers and statistics to extract information (patterns, entities, topics, etc) from text
- We typically use text to make inferences about some latent variable
    - Observe transcripts, news articles, social media posts, etc.
    - Make inferences about things we can't directly observe like ideology
- Or to categorize texts into different classes

Text Analysis

- Text analysis uses computers and statistics to extract information (patterns, entities, topics, etc) from text
- We typically use text to make inferences about some latent variable
    - Observe transcripts, news articles, social media posts, etc.
    - Make inferences about things we can't directly observe like ideology
- Or to categorize texts into different classes
- But, new ways to use them in experiments, etc.!

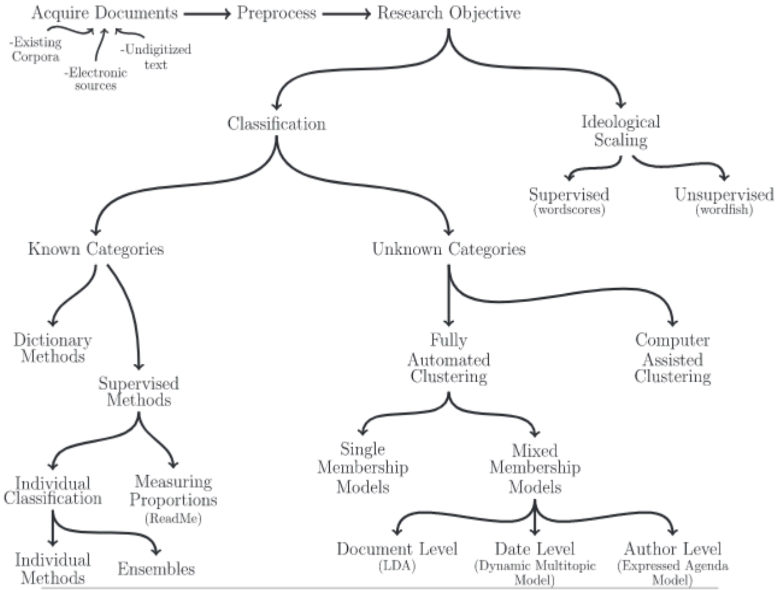


Fig. 1 An overview of text as data methods.

Supervised vs Unsupervised

- Supervised methods

Supervised vs Unsupervised

- Supervised methods
    - Explore the relationship between inputs and a labeled set of outputs

Supervised vs Unsupervised

- Supervised methods
    - Explore the relationship between inputs and a labeled set of outputs
- Unsupervised methods

Supervised vs Unsupervised

- Supervised methods
    - Explore the relationship between inputs and a labeled set of outputs
- Unsupervised methods
    - Explore the hidden or latent structure in unlabeled data

Assumption

Bag of Words

- We typically format text data as a document term matrix
    - Rows = documents, columns = terms
- Disregards grammar and word order
- *Terms* are typically single words, but can be other things that we *tokenize* the text into
    - *Tokens* are just units of the text used in the analysis
    - We can tokenize text into sentences, paragraphs, n-grams (collections of words or sentences), etc.

	I	love	programming	in	R
Document1	0	2	3	0	3
Document2	1	1	1	1	1
Document3	1	1	0	3	0

Classifying Known Categories

Dictionary Methods

These methods typically-

- classify documents into known categories

Dictionary Methods

These methods typically-

- classify documents into known categories
  - *This is an angry tweet*

Dictionary Methods

These methods typically-

- classify documents into known categories
  - *This is an angry tweet*
- Assess how associated a document is to a category

Dictionary Methods

These methods typically-

- classify documents into known categories
  - *This is an angry tweet*
- Assess how associated a document is to a category
  - *This is* sort of *an angry tweet*

Dictionary Methods

These methods typically-

- classify documents into known categories
  - *This is an angry tweet*
- Assess how associated a document is to a category
  - *This is* sort of *an angry tweet*
- Dictionary methods use sets of words that are associated with certain labels to do this

Dictionary Methods

- *Dictionaries* are sets of words with associated labels

Dictionary Methods

- *Dictionaries* are sets of words with associated labels
  - Angry = [Hate, Irritate]

Dictionary Methods

- *Dictionaries* are sets of words with associated labels
    - Angry = [Hate, Irritate]
    - Happy = [Ecstatic, Pleased]

Dictionary Methods

- *Dictionaries* are sets of words with associated labels
  - Angry = [Hate, Irritate]
  - Happy = [Ecstatic, Pleased]
- Dictionary models classify documents using the score of each word in the document, the number of word occurrences, and the number of total words

Dictionary Methods

- A simple example:

Dictionary Methods

- A simple example:
  - Positive (+1) = [Great, Fun, Amazing, Enjoyable]

Dictionary Methods

- A simple example:
  - Positive (+1) = [Great, Fun, Amazing, Enjoyable]
  - Negative (-1) = [Sour, Sad, Boring, Lame]

Dictionary Methods

- A simple example:
  - Positive (+1) = [Great, Fun, Amazing, Enjoyable]
  - Negative (-1) = [Sour, Sad, Boring, Lame]
  - Document = The ride was amazing and fun, but the snacks were pretty boring.

Dictionary Methods

- A simple example:
  - Positive (+1) = [Great, Fun, Amazing, Enjoyable]
  - Negative (-1) = [Sour, Sad, Boring, Lame]
  - Document = The ride was amazing and fun, but the snacks were pretty boring.
  - 2 positive words, 1 negative word

Dictionary Methods

- A simple example:
  - Positive (+1) = [Great, Fun, Amazing, Enjoyable]
  - Negative (-1) = [Sour, Sad, Boring, Lame]
  - Document = The ride was amazing and fun, but the snacks were pretty boring.
  - 2 positive words, 1 negative word
  - $Score = (Number_{pos} * Score_{pos} + Number_{neg} * Score_{neg}) / N_{words}$

Dictionary Methods

- A simple example:
  - Positive (+1) = [Great, Fun, Amazing, Enjoyable]
  - Negative (-1) = [Sour, Sad, Boring, Lame]
  - Document = The ride was amazing and fun, but the snacks were pretty boring.
  - 2 positive words, 1 negative word
  - Score = ($Number_{pos}$ * $Score_{pos}$ + $Number_{neg}$ * $Score_{neg}$) / N_{words}
  - Score = (2 * 1 + 1 * - 1)

Dictionary Methods

- A simple example:
  - Positive (+1) = [Great, Fun, Amazing, Enjoyable]
  - Negative (-1) = [Sour, Sad, Boring, Lame]
  - Document = The ride was amazing and fun, but the snacks were pretty boring.
  - 2 positive words, 1 negative word
  - Score = $(\text{Number}_{pos} * \text{Score}_{pos} + \text{Number}_{neg} * \text{Score}_{neg}) \,/\, \text{N}_{words}$
  - Score = $(2 * 1 + 1 * -1)$
  - Score = 0.3

Dictionary Methods

- The words in the dictionary matter a lot

Dictionary Methods

- The words in the dictionary matter a lot
- Lots of general-purpose dictionaries

Dictionary Methods

- The words in the dictionary matter a lot
- Lots of general-purpose dictionaries
  - But these often struggle when applied to niche sets of documents or new domains

Dictionary Methods

- The words in the dictionary matter a lot
- Lots of general-purpose dictionaries
    - But these often struggle when applied to niche sets of documents or new domains
    - E.g. scoring the sentiment of Islamic State propaganda without a purpose-built dictionary

Dictionary Methods

- The words in the dictionary matter a lot
- Lots of general-purpose dictionaries
    - But these often struggle when applied to niche sets of documents or new domains
    - E.g. scoring the sentiment of Islamic State propaganda without a purpose-built dictionary
        - Words like *martyr* and *caliphate*

Dictionary Methods

- The words in the dictionary matter a lot
- Lots of general-purpose dictionaries
  - But these often struggle when applied to niche sets of documents or new domains
  - E.g. scoring the sentiment of Islamic State propaganda without a purpose-built dictionary
    - Words like *martyr* and *caliphate*
  - Easy to create custom dictionaries or add to existing dictionaries

Supervised Learning

- Also classifies known categories

Supervised Learning

- Also classifies known categories
- Generalizes the intuition of dictionary methods

Supervised Learning

- Also classifies known categories
- Generalizes the intuition of dictionary methods
  - Now the model learns from the data how features are associated with different categories

Supervised Learning

- Also classifies known categories
- Generalizes the intuition of dictionary methods
    - Now the model learns from the data how features are associated with different categories
    - Uses a labeled subset of the data to learn this relationship

Supervised Learning

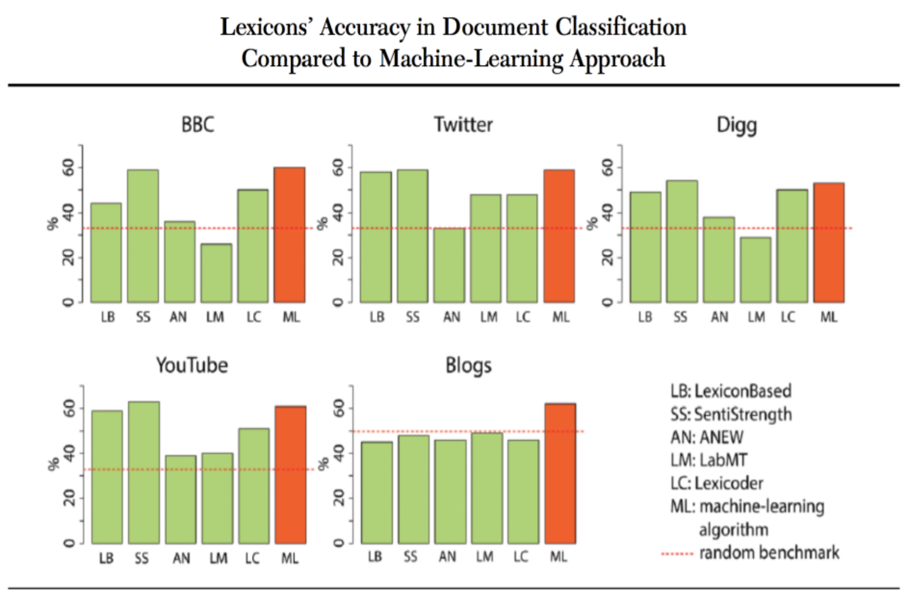
- Also classifies known categories
- Generalizes the intuition of dictionary methods
  - Now the model learns from the data how features are associated with different categories
  - Uses a labeled subset of the data to learn this relationship
    - *Positive*: The ocean is so relaxing

Supervised Learning

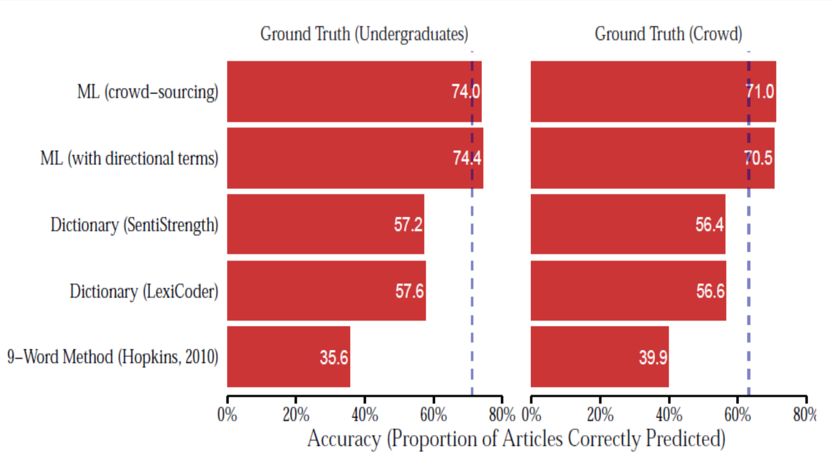
- Also classifies known categories
- Generalizes the intuition of dictionary methods
  - Now the model learns from the data how features are associated with different categories
  - Uses a labeled subset of the data to learn this relationship
    - *Positive*: The ocean is so relaxing
    - *Negative*: The ocean makes me sick

Supervised Learning

- Also classifies known categories
- Generalizes the intuition of dictionary methods
  - Now the model learns from the data how features are associated with different categories
  - Uses a labeled subset of the data to learn this relationship
    - *Positive*: The ocean is so relaxing
    - *Negative*: The ocean makes me sick
- Supervised learning will outperform dictionary methods in classification tasks given sufficient training set



Gonzalez-Bailon and Paltoglu (2015)



Barbera et al (2017)

Supervised Learning

- How do we label documents?

Supervised Learning

- How do we label documents?
  - Some projects rely on experts (e.g. Manifesto Project Database)

Supervised Learning

- How do we label documents?
  - Some projects rely on experts (e.g. Manifesto Project Database)
  - But we can label documents ourselves or with RAs

Supervised Learning

- How do we label documents?
    - Some projects rely on experts (e.g. Manifesto Project Database)
    - But we can label documents ourselves or with RAs
    - Crowdsourcing on platforms like MTurk is a cheap/effective option as well

Supervised Learning

- How do we label documents?
  - Some projects rely on experts (e.g. Manifesto Project Database)
  - But we can label documents ourselves or with RAs
  - Crowdsourcing on platforms like MTurk is a cheap/effective option as well
  - We don't necessarily care about the sample of coders, just how how well they can label documents

Supervised Learning

- How do we label documents?
  - Some projects rely on experts (e.g. Manifesto Project Database)
  - But we can label documents ourselves or with RAs
  - Crowdsourcing on platforms like MTurk is a cheap/effective option as well
  - We don't necessarily care about the sample of coders, just how how well they can label documents
  - We do want to use a number of tools to assess the performance of the labels (percent agreement, correlation, etc.)

Supervised Learning

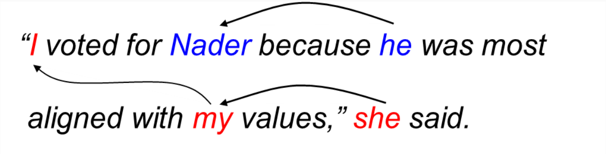
- How do you know what documents to label/classify?

Supervised Learning

- How do you know what documents to label/classify?
- You might care about mentions of a politician, but matching their name excludes references like *they*

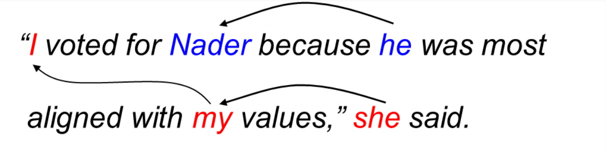
Supervised Learning

- How do you know what documents to label/classify?
- You might care about mentions of a politician, but matching their name excludes references like *they*
- Consider using coreference resolution



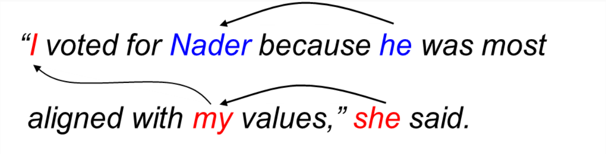
Supervised Learning

- How do you know what documents to label/classify?
- You might care about mentions of a politician, but matching their name excludes references like *they*
- Consider using coreference resolution
  - Locates all words that refer to the same entity in a text



Supervised Learning

- How do you know what documents to label/classify?
- You might care about mentions of a politician, but matching their name excludes references like *they*
- Consider using coreference resolution
    - Locates all words that refer to the same entity in a text



- Stanford CoreNLP natural language software

Supervised Learning

- Lots of classifiers used in this setup

Supervised Learning

- Lots of classifiers used in this setup
  - SVM, Random Forest, Naive Bayes, etc.

Supervised Learning

- Lots of classifiers used in this setup
  - SVM, Random Forest, Naive Bayes, etc.
- Typical need to cross validate, etc.

Supervised Learning

- Lots of classifiers used in this setup
  - SVM, Random Forest, Naive Bayes, etc.
- Typical need to cross validate, etc.
- Need to balance the bias-variance tradeoff

Supervised Learning

- Lots of classifiers used in this setup
  - SVM, Random Forest, Naive Bayes, etc.
- Typical need to cross validate, etc.
- Need to balance the bias-variance tradeoff
  - Overfit to training set, poor performance in test set (no bias, creates variance)

Supervised Learning

- Lots of classifiers used in this setup
  - SVM, Random Forest, Naive Bayes, etc.
- Typical need to cross validate, etc.
- Need to balance the bias-variance tradeoff
  - Overfit to training set, poor performance in test set (no bias, creates variance)
  - Don't overfit and accept bias but reduce variance (allow noise in training set)

Supervised Learning - Greene and Lucas (2020)

- *Question*: How can we measure relationships between terrorists/rebel groups?

Supervised Learning - Greene and Lucas (2020)

- *Question*: How can we measure relationships between terrorists/rebel groups?
- *Argument*: Use what they say about each other in their official publications

Supervised Learning - Greene and Lucas (2020)

- *Question*: How can we measure relationships between terrorists/rebel groups?
- *Argument*: Use what they say about each other in their official publications
- *Documents*: sentences that mention another group

Supervised Learning - Greene and Lucas (2020)

- *Question*: How can we measure relationships between terrorists/rebel groups?
- *Argument*: Use what they say about each other in their official publications
- *Documents*: sentences that mention another group
- *Main model*: Classified documents using an SVM

Supervised Learning - Greene and Lucas (2020)

- *Question*: How can we measure relationships between terrorists/rebel groups?
- *Argument*: Use what they say about each other in their official publications
- *Documents*: sentences that mention another group
- *Main model*: Classified documents using an SVM
- *Finding*: Monthly/yearly aggregated sentiment tracks real-world relationships

Supervised Learning - Greene and Lucas (2020)

- *Question*: How can we measure relationships between terrorists/rebel groups?
- *Argument*: Use what they say about each other in their official publications
- *Documents*: sentences that mention another group
- *Main model*: Classified documents using an SVM
- *Finding*: Monthly/yearly aggregated sentiment tracks real-world relationships
    - Measure can be used in future observational studies

Classifying Unknown Categories

Topic Models

- Algorithms that find *themes* in text

Topic Models

- Algorithms that find *themes* in text
- No prior information, no training set, no human input at all (generally) except the number of topics

Topic Models

- Algorithms that find *themes* in text
- No prior information, no training set, no human input at all (generally) except the number of topics
- LDA is most common model

Topic Models

- Algorithms that find *themes* in text
- No prior information, no training set, no human input at all (generally) except the number of topics
- LDA is most common model
    - Describes how the documents in the corpus were created

Topic Models

- Algorithms that find *themes* in text
- No prior information, no training set, no human input at all (generally) except the number of topics
- LDA is most common model
    - Describes how the documents in the corpus were created
    - Assume some number of topics

Topic Models

- Algorithms that find *themes* in text
- No prior information, no training set, no human input at all (generally) except the number of topics
- LDA is most common model
  - Describes how the documents in the corpus were created
  - Assume some number of topics
  - Each topic is a distribution over words

Topic Models

- Algorithms that find *themes* in text
- No prior information, no training set, no human input at all (generally) except the number of topics
- LDA is most common model
  - Describes how the documents in the corpus were created
  - Assume some number of topics
  - Each topic is a distribution over words
  - Each document is comprised of words generated by a multinomial distribution (one for each topic)

Topic Models

- Every document is a mixture of topics

Topic Models

- Every document is a mixture of topics
- Every topic is a mixture of words

Topic Models

- Every document is a mixture of topics
- Every topic is a mixture of words
- Social scientists often use topic modeling for exploratory and descriptive analyses

Topic Models

- Every document is a mixture of topics
- Every topic is a mixture of words
- Social scientists often use topic modeling for exploratory and descriptive analyses
- Plenty of other applications though

Topic Models

- Every document is a mixture of topics
- Every topic is a mixture of words
- Social scientists often use topic modeling for exploratory and descriptive analyses
- Plenty of other applications though
- Extensions of TMs employ document metadata (author information, context, time, etc) in a regression framework

Topic Models

- Every document is a mixture of topics
- Every topic is a mixture of words
- Social scientists often use topic modeling for exploratory and descriptive analyses
- Plenty of other applications though
- Extensions of TMs employ document metadata (author information, context, time, etc) in a regression framework
    - Applications to experiments and observational studies

Topic Models - Terman (2017)

- *Question*: How are Muslim women portrayed in the US news media?

Topic Models - Terman (2017)

- *Question*: How are Muslim women portrayed in the US news media?
- *Argument*: Coverage is driven by Orientalism and narratives of oppression

Topic Models - Terman (2017)

- *Question*: How are Muslim women portrayed in the US news media?
- *Argument*: Coverage is driven by Orientalism and narratives of oppression
- *Documents*: New York Times and Washington Post articles

Topic Models - Terman (2017)

- *Question*: How are Muslim women portrayed in the US news media?
- *Argument*: Coverage is driven by Orientalism and narratives of oppression
- *Documents*: New York Times and Washington Post articles
- *Main model*: Structural Topic Model (integrates metadata and TMs with a regression framework)

Topic Models - Terman (2017)

- *Question*: How are Muslim women portrayed in the US news media?
- *Argument*: Coverage is driven by Orientalism and narratives of oppression
- *Documents*: New York Times and Washington Post articles
- *Main model*: Structural Topic Model (integrates metadata and TMs with a regression framework)
    - Topics = Region + Year Control

Topic Models - Terman (2017)

- *Question*: How are Muslim women portrayed in the US news media?
- *Argument*: Coverage is driven by Orientalism and narratives of oppression
- *Documents*: New York Times and Washington Post articles
- *Main model*: Structural Topic Model (integrates metadata and TMs with a regression framework)
    - Topics = Region + Year Control
- *Finding*: See figure

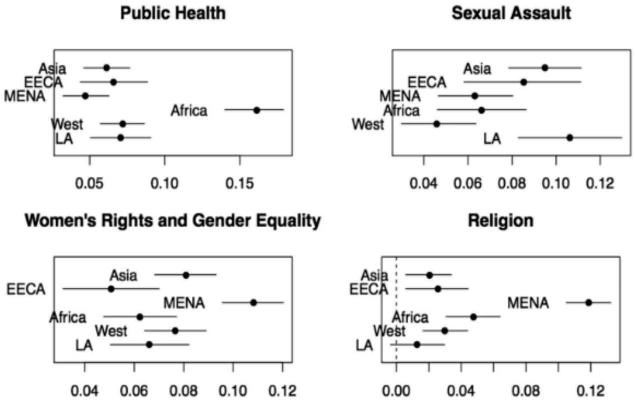


Figure 4. Expected document proportions for four topics

- Expected document proportion of an unseen document as a function of region and year
- Terman (2017)

R sesh!

R sesh!

https://github.com/caleblucas/text_analysis