Lecture 16: Bag of Words and More

Naijia Liu

March 26 2024

Final Project Poster

- More info on course website: poster samples and resources.
- Gov 51 final poster session will happen on 4/23 Tuesday usual class time, with light refreshment
- Workshop on 4/4 (attendance is required, contents are optional)

  I will record myself for a 40 min coding session.

  James will give a lecture on regression discontinuity in person.

  RD is widely applied and will be super super super helpful for final project and / or thesis.

A New Source of Data

- Measuring agenda setting in interactive political communication (Rossiter, 2022)

A New Source of Data

- Measuring agenda setting in interactive political communication (Rossiter, 2022)
- Within interactions, such as debates, deliberations, and discussions, actors can set the agenda by shifting others' attention to their preferred topics.

  With non-text data, we can (maybe) study it by variables such as speaking duration in a debate, votes and etc.

A New Source of Data

- Measuring agenda setting in interactive political communication (Rossiter, 2022)
- Within interactions, such as debates, deliberations, and discussions, actors can set the agenda by shifting others' attention to their preferred topics.

  With non-text data, we can (maybe) study it by variables such as speaking duration in a debate, votes and etc.
- By analyzing the transcript of debates, we can locate where topic shifts occur within an interaction in order to measure the relative agenda-setting power of actors.

A New Source of Data

- Measuring agenda setting in interactive political communication (Rossiter, 2022)
- Within interactions, such as debates, deliberations, and discussions, actors can set the agenda by shifting others' attention to their preferred topics.

  With non-text data, we can (maybe) study it by variables such as speaking duration in a debate, votes and etc.
- By analyzing the transcript of debates, we can locate where topic shifts occur within an interaction in order to measure the relative agenda-setting power of actors.
- Successfully setting the agenda can shape an interaction's outcomes.

Example from 2016 Presidential Debate

Holt: We are at—we are at the final question.

Clinton: Well, one thing. One thing, Lester.

Holt: Very quickly, because we're at the final question now.

Clinton: You know, he tried to switch from looks to stamina. But this is a man who has called women pigs, slobs and dogs, and someone who has said pregnancy is an inconvenience to employers, who has said...

Intuition Behind the Method

- Data generating process of texts:

  We have a topic in mind, which determines the probability distribution of vocabularies. And to speak, we randomly draw words from the vocabularies with such probabilities.

Intuition Behind the Method

- Data generating process of texts:

  We have a topic in mind, which determines the probability distribution of vocabularies. And to speak, we randomly draw words from the vocabularies with such probabilities.

  - Food: Pasta (high probability), Coke (high probability), Rain, Good, Bad, Temperature, Warm.

Intuition Behind the Method

- Data generating process of texts:

  We have a topic in mind, which determines the probability distribution of vocabularies. And to speak, we randomly draw words from the vocabularies with such probabilities.

  - Food: Pasta (high probability), Coke (high probability), Rain, Good, Bad, Temperature, Warm.
  - Weather: Pasta, Coke, Rain (high probability), Good, Bad, Temperature (high probability), Warm.

Intuition Behind the Method

- Data generating process of texts:

  We have a topic in mind, which determines the probability distribution of vocabularies. And to speak, we randomly draw words from the vocabularies with such probabilities.

  - Food: Pasta (high probability), Coke (high probability), Rain, Good, Bad, Temperature, Warm.
  - Weather: Pasta, Coke, Rain (high probability), Good, Bad, Temperature (high probability), Warm.

- If we believe in such generating process, we will be able to back off each topics from the transcripts.

Intuition Behind the Method

- Data generating process of texts:

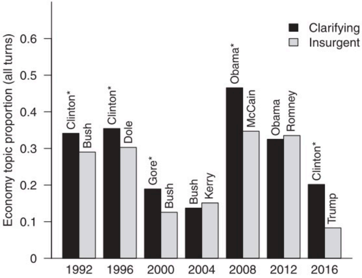
  We have a topic in mind, which determines the probability distribution of vocabularies. And to speak, we randomly draw words from the vocabularies with such probabilities.

  - Food: Pasta (high probability), Coke (high probability), Rain, Good, Bad, Temperature, Warm.
  - Weather: Pasta, Coke, Rain (high probability), Good, Bad, Temperature (high probability), Warm.
- If we believe in such generating process, we will be able to back off each topics from the transcripts.
- Then, author was able to measure topic changes throughout the document.

Candidates Behave Differently

Clarifying candidates tend to switch to economic topics.

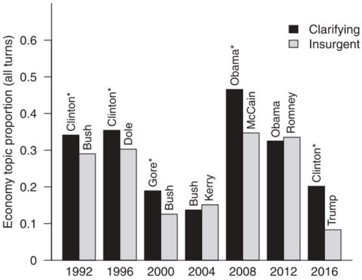
Figure: Left: All turns; Right: Topic changing turns

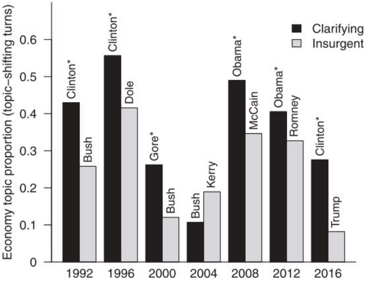


Candidates Behave Differently

Clarifying candidates tend to switch to economic topics.

Figure: Left: All turns; Right: Topic changing turns





Does public opinion affect political speech?

- Does public opinion affect: (Hager and Hilbig, 2020, AJPS)

Does public opinion affect political speech?

- Does public opinion affect: (Hager and Hilbig, 2020, AJPS)
  - what topics politicians address

Does public opinion affect political speech?

- Does public opinion affect: (Hager and Hilbig, 2020, AJPS)
  - what topics politicians address
  - what positions they endorse

Does public opinion affect political speech?

- Does public opinion affect: (Hager and Hilbig, 2020, AJPS)
    - what topics politicians address
    - what positions they endorse
- German government declassified public opinion research to its cabinet members.

Does public opinion affect political speech?

- Does public opinion affect: (Hager and Hilbig, 2020, AJPS)
  - what topics politicians address
  - what positions they endorse
- German government declassified public opinion research to its cabinet members.
- **Linguistic similarity** as a measure of congruence

Does public opinion affect political speech?

- Does public opinion affect: (Hager and Hilbig, 2020, AJPS)
  - what topics politicians address
  - what positions they endorse
- German government declassified public opinion research to its cabinet members.
- **Linguistic similarity** as a measure of congruence
- Exposure to public opinion research leads politicians to markedly change their speech

Does public opinion affect political speech?

- Cosine similarity to measure linguistic similarity.
  - We will support labor unions.
  - Labor unions should be supported.
- Model:

$$\text{Cosine Sim} = \beta_0 + \beta_1 \cdot \text{Exposure} + \beta X + \epsilon$$

- More details + a very smart RD design. Take a look at the paper if interested!

Exposure Leads to Higher Similarity

Figure: Speeches follow public opinion.

	Cosine Similarity	
	(1)	(2)
Exposure	0.0137**	0.0128**
	(0.0066)	(0.0057)
Covariates	No	Yes
Observations	5,684	5,684
Mean of DV	0.1263	
SD of DV	0.0976	
Effect size in SD	0.1413	0.1319

The Bag-of-Words Model

- To achieve above studies, we need to transform text data into something simpler.

The Bag-of-Words Model

- To achieve above studies, we need to transform text data into something simpler.
- The bag-of-words model is a simple and widely used approach to analyze textual data

The Bag-of-Words Model

- To achieve above studies, we need to transform text data into something simpler.
- The bag-of-words model is a simple and widely used approach to analyze textual data
- The bag-of-words model represents a text as a collection of words, ignoring the order and structure of the sentences

The Bag-of-Words Model

- To achieve above studies, we need to transform text data into something simpler.
- The bag-of-words model is a simple and widely used approach to analyze textual data
- The bag-of-words model represents a text as a collection of words, ignoring the order and structure of the sentences
- Assumption: the frequency of words in a text can provide valuable information about the content of the text

How the Bag-of-Words Model Works

- The bag-of-words model involves the following steps:

How the Bag-of-Words Model Works

- The bag-of-words model involves the following steps:
    1. Tokenization: dividing a text into individual words or tokens

How the Bag-of-Words Model Works

- The bag-of-words model involves the following steps:
  1. Tokenization: dividing a text into individual words or tokens
  2. Counting: counting the frequency of each word in the text

How the Bag-of-Words Model Works

- The bag-of-words model involves the following steps:
  1. Tokenization: dividing a text into individual words or tokens
  2. Counting: counting the frequency of each word in the text
  3. Vectorization: representing the text as a vector of word frequencies

Example of Bag-of-Words Model

- Suppose we have the following two sentences:
  - Sentence 1: The great fox loves the lazy dog
  - Sentence 2: The lazy dog sleeps all day

Example of Bag-of-Words Model

- Suppose we have the following two sentences:
  - Sentence 1: The great fox loves the lazy dog
  - Sentence 2: The lazy dog sleeps all day
- The bag-of-words representation of these two sentences would be:

	the	great	fox	loves	lazy	dog	sleeps	all	day
1	2	1	1	1	1	1	0	0	0
2	1	0	0	0	1	1	1	1	1

Example of Bag-of-Words Model

- Suppose we have the following two sentences:
  - Sentence 1: The great fox loves the lazy dog
  - Sentence 2: The lazy dog sleeps all day
- The bag-of-words representation of these two sentences would be:

	the	great	fox	loves	lazy	dog	sleeps	all	day
1	2	1	1	1	1	1	0	0	0
2	1	0	0	0	1	1	1	1	1

- End up with a N (number of documents) by P (unique vocabulary) document term matrix.

Applications of the Bag-of-Words Model

- The bag-of-words model can be used for a variety of applications such as

Applications of the Bag-of-Words Model

- The bag-of-words model can be used for a variety of applications such as
  - Topic modeling: identifying the topics or themes present in a collection of texts

Applications of the Bag-of-Words Model

- The bag-of-words model can be used for a variety of applications such as
  - Topic modeling: identifying the topics or themes present in a collection of texts
  - Sentiment analysis: determining the sentiment of a text, such as positive or negative

Applications of the Bag-of-Words Model

- The bag-of-words model can be used for a variety of applications such as
  - Topic modeling: identifying the topics or themes present in a collection of texts
  - Sentiment analysis: determining the sentiment of a text, such as positive or negative
  - Text classification: classifying a text into predefined categories based on its content

Sentiment Analysis: Dictionary Method

Table: Text Data

	the	great	fox	loves	lazy	dog	sleeps	all	day
1	2	1	1	1	1	1	0	0	0
2	1	0	0	0	1	1	1	1	1

- Imagine a dictionary with following words and labels.

Sentiment Analysis: Dictionary Method

Table: Text Data

	the	great	fox	loves	lazy	dog	sleeps	all	day
1	2	1	1	1	1	1	0	0	0
2	1	0	0	0	1	1	1	1	1

- Imagine a dictionary with following words and labels.
  - Positive: great, love

Sentiment Analysis: Dictionary Method

Table: Text Data

	the	great	fox	loves	lazy	dog	sleeps	all	day
1	2	1	1	1	1	1	0	0	0
2	1	0	0	0	1	1	1	1	1

- Imagine a dictionary with following words and labels.
  - Positive: great, love
  - Negative: lazy

Sentiment Analysis: Dictionary Method

Table: Text Data

	the	great	fox	loves	lazy	dog	sleeps	all	day
1	2	1	1	1	1	1	0	0	0
2	1	0	0	0	1	1	1	1	1

- Imagine a dictionary with following words and labels.
  - Positive: great, love
  - Negative: lazy
  - Neutral: rest of the words.

Sentiment Analysis: Dictionary Method

Table: Text Data

	the	great	fox	loves	lazy	dog	sleeps	all	day
1	2	1	1	1	1	1	0	0	0
2	1	0	0	0	1	1	1	1	1

- Imagine a dictionary with following words and labels.
  - Positive: great, love
  - Negative: lazy
  - Neutral: rest of the words.
- We can calculate sentiment score for each sentence above.

Sentiment Analysis: Dictionary Method

Table: Text Data

	the	great	fox	loves	lazy	dog	sleeps	all	day
1	2	1	1	1	1	1	0	0	0
2	1	0	0	0	1	1	1	1	1

- Sentence 1: $1 * 1 + 1 * 1 + (-1) * 1 = 1$
- Sentence 2: $(-1) * 1 = -1$

Sentiment Analysis: Dictionary Method

- Sentence 1 is more positive than sentence 2.

Sentiment Analysis: Dictionary Method

- Sentence 1 is more positive than sentence 2.
- Dictionary method is cheap and fast. (more to come on Thursday)

Sentiment Analysis: Dictionary Method

- Sentence 1 is more positive than sentence 2.
- Dictionary method is cheap and fast. (more to come on Thursday)
- Weighing every word equally.

  unimpressive, bad, terrible, bizarre.

Sentiment Analysis: Dictionary Method

- Sentence 1 is more positive than sentence 2.
- Dictionary method is cheap and fast. (more to come on Thursday)
- Weighing every word equally.

  unimpressive, bad, terrible, bizarre.
- We lose the order and structure of sentences.

  It is not bad.

  $\rightsquigarrow$ will have a negative sentiment score!

Sentiment Analysis: Supervised Learning

Table: Text Data

	the	great	fox	loves	lazy	dog	sleeps	all	day
1	2	1	1	1	1	1	0	0	0
2	1	0	0	0	1	1	1	1	1

- Dictionary method has its limitations, hence we want to bring in human coders!
- Coders will be able to label the sentences for us:
  - Sentence 1: Positive
  - Sentence 2: Negative

Sentiment Analysis: Supervised Learning

- We use this info to train a prediction model:

$$\text{Sentiment} = \beta_0 + \beta_1 \cdot \text{the} + \beta_1 \cdot \text{great} + \beta_1 \cdot \text{fox} + \cdots + \epsilon$$

Sentiment Analysis: Supervised Learning

- We use this info to train a prediction model:

$$\text{Sentiment} = \beta_0 + \beta_1 \cdot \text{the} + \beta_1 \cdot \text{great} + \beta_1 \cdot \text{fox} + \cdots + \epsilon$$

- We input the count of each word as the value of the variable.

Sentiment Analysis: Supervised Learning

- We use this info to train a prediction model:

$$\text{Sentiment} = \beta_0 + \beta_1 \cdot \text{the} + \beta_1 \cdot \text{great} + \beta_1 \cdot \text{fox} + \cdots + \epsilon$$

- We input the count of each word as the value of the variable.
- Say now we have a new sentence: The lazy dog loves his owner.

Sentiment Analysis: Supervised Learning

- We use this info to train a prediction model:

$$\text{Sentiment} = \beta_0 + \beta_1 \cdot \text{the} + \beta_1 \cdot \text{great} + \beta_1 \cdot \text{fox} + \cdots + \epsilon$$

- We input the count of each word as the value of the variable.
- Say now we have a new sentence: The lazy dog loves his owner.
  - First, we tokenize and vectorize it:

	the	great	fox	loves	lazy	dog	sleeps	all	day
1	1	0	0	1	1	1	0	0	0

Sentiment Analysis: Supervised Learning

- We use this info to train a prediction model:

$$\text{Sentiment} = \beta_0 + \beta_1 \cdot \text{the} + \beta_1 \cdot \text{great} + \beta_1 \cdot \text{fox} + \cdots + \epsilon$$

- We input the count of each word as the value of the variable.
- Say now we have a new sentence: The lazy dog loves his owner.
  - First, we tokenize and vectorize it:

	the	great	fox	loves	lazy	dog	sleeps	all	day
1	1	0	0	1	1	1	0	0	0

  - Then, we use the trained model to predict the sentiment of it by plugging in values for the variables.

Sentiment Analysis: Supervised Learning

- Real world text data contains thousands of unique vocabularies.

Sentiment Analysis: Supervised Learning

- Real world text data contains thousands of unique vocabularies.
- Some of the words are not useful in predicting sentiments.

  $\rightsquigarrow$ the, is, and ....

Sentiment Analysis: Supervised Learning

- Real world text data contains thousands of unique vocabularies.
- Some of the words are not useful in predicting sentiments.

  $\rightsquigarrow$ the, is, and ....
- We can use Lasso to select variables for the following model:

$$\text{Sentiment} = \beta_0 + \beta_1 \cdot \text{the} + \beta_1 \cdot \text{great} + \beta_1 \cdot \text{fox} + \cdots + \epsilon$$

Sentiment Analysis: Supervised Learning

- Real world text data contains thousands of unique vocabularies.
- Some of the words are not useful in predicting sentiments.

  $\rightsquigarrow$ the, is, and ....
- We can use Lasso to select variables for the following model:

$$\text{Sentiment} = \beta_0 + \beta_1 \cdot \text{the} + \beta_1 \cdot \text{great} + \beta_1 \cdot \text{fox} + \cdots + \epsilon$$

- Lasso will select for us the variables (vocabularies) with a substantively large enough coefficient in predicting the sentiment.

Sentiment Analysis: Supervised Learning

- Human coders understand better the context of the words.
- Supervised learning is more costly and slower.
- Models cannot work with new vocabularies that are not covered in the training data.
  - The lazy dog loves **his owner**.

Limitations of the Bag-of-Words Model

- The bag-of-words model has several limitations, including:
  - It ignores the order and structure of words in a sentence, which can result in the loss of important information
  - It treats all words as equally important, even though some words may be more informative than others
  - It does not capture the meaning of words, only their frequency in a text
- Despite these limitations, the bag-of-words model is still a useful and widely used approach to analyze textual data

N-gram Tokenization

- We want to preserve the order and structure better.

N-gram Tokenization

- We want to preserve the order and structure better.
- Take bi-gram as an example:

  It is not very bad.

  $\rightsquigarrow$ It is, is not, not very, very bad.

N-gram Tokenization

- We want to preserve the order and structure better.
- Take bi-gram as an example:

  It is not very bad.

  $\rightsquigarrow$ It is, is not, not very, very bad.
- Then let's go with tri-gram:

  It is not very bad.

  $\rightsquigarrow$ It is not, is not very, not very bad.

Common Practices

- Pre-processing
  - Get rid of the most / least frequent words.
  - Stemming of the words.
- Pre-processing decisions have profound effects on the results of real models for real data. (Denny and Spirling, 2018, Political Analysis)

Bag-of-Words Model

- Vectorization of words is the foundation of all most all text analysis methods.
- We will try a simple text analysis together on Thursday.
- We will discuss un-supervised learning next week.

  We don't have an outcome variable of interest, but just to summarize the text data.