Analyzing Bank Cybersecurity Disclosures Using Machine Learning

Simpson Zhang

Discussion by Minchul Shin[1] (FRB Philadelphia)

Interagency Risk Quantification Forum
November 29–30, 2022

[1] **Disclaimer:** The views expressed here are my own and do not necessarily represent the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System.

This paper ...

Estimate Latent Dirichlet Allocation (LDA) model

- Use two sets of annual reports by Verizon and Ponemon
- Treat each paragraph from the report as a “document”

Use estimated LDA model to evaluate discussions on cybersecurity from 10-K reports (50 large U.S. banks for 10 years)

- Consider three “report quality metrics”: Length score (LS), Match score (MS), Product score (PS)

Study behavior of those computed metrics over time and over cross-section

Plan for my discussion

1. Review LDA
2. Two comments
   - Comment 1: Theoretical justification for proposed metrics
   - Comment 2: Validation for internal consistency

Review of
Latent Dirichlet Allocation (LDA)

LDA is one of the most popular topic modeling methods that classify documents into topics

LDA – Data generating mechanism

There are **V** words ever appeared in all documents

$$\{word_1, word_2, word_3, ..., word_V\}$$

There are **N** common topics

$$\{topic_1, topic_2, topic_3, ..., topic_N\}$$

There are **M** documents; *i*-th document is a sequence of $\boldsymbol{N_i}$ words

$$Document_i = \{w_{i,1}, w_{i,2}, w_{i,3}, ..., w_{i,N_i}\}$$

Each word in the document is chosen in a probabilistic way

LDA – Data generating mechanism

For i-th document, topics are assigned based on **topic distribution**

$$\theta_i = (\theta_{1i}, \theta_{2i}, ..., \theta_{Ki}) \in \Delta^{K-1}$$

For j-th topic, words are generated based on **word distribution**

$$\phi_j = (\phi_{1j}, \phi_{2j}, ..., \phi_{Vj}) \in \Delta^{V-1}$$

k-th word in the i-th document is generated by

1. Choose topic $z_{ik} = j$ where $j \sim Multinomial(\theta_i)$
2. Choose word $w_{ik} = w$ where $w \sim Multinomial(\phi_{z_{ik}})$

LDA – Example 1

Consider a small vocabulary ($V = 5$)

$$\mathcal{V} = \{\textit{credentials}, \textit{breach}, \textit{this}, \textit{maybe}, \textit{stolen}\}$$

Suppose there are two topics with equal probability, $\theta = \{0.5, 0.5\}$

Each topic assigns probability of word appearance in the vocabulary,

$$\text{Topic 1}: \ \phi_1 = \{0.3, 0.4, 0.0, 0.0, 0.3\}$$
$$\text{Topic 2}: \ \phi_2 = \{0.0, 0.0, 0.5, 0.5, 0.0\}$$

- The first topic assigns probability on “credentials”, “breach”, “stolen”.
- The second topic assigns probability on “this” and “maybe”

LDA example 2-1

Suppose there are two sentences in *i*-th document. LDA starts with empty boxes to be filled in:

LDA example 2-2

Suppose there are two sentences in *i*-th document. LDA starts with empty boxes to be filled in:

***		

For the first cell,

- Draw a topic with probability θ_i. Suppose the first topic is selected
- Draw a word from $\mathcal{V}$ with probability ϕ_1. Suppose "credentials" is chosen

where

$$\mathcal{V} = \{\textit{credentials}, \textit{breach}, \textit{this}, \textit{maybe}, \textit{stolen}\}$$
$$\phi_1 = \{0.3, 0.4, 0.0, 0.0, 0.3\}$$
$$\phi_2 = \{0.0, 0.0, 0.5, 0.5, 0.0\}$$
$$\theta = (0.5, 0.5)$$

LDA example 2-3

Suppose there are two sentences in *i*-th document. LDA starts with empty boxes to be filled in:

Credentials		

For the first cell,

- Draw a topic with probability θ_i. Suppose the first topic is selected
- Draw a word from $\mathcal{V}$ with probability ϕ_1. Suppose "credentials" is chosen

where

$$\mathcal{V} = \{\textit{credentials}, \textit{breach}, \textit{this}, \textit{maybe}, \textit{stolen}\}$$
$$\phi_1 = \{0.3, 0.4, 0.0, 0.0, 0.3\}$$
$$\phi_2 = \{0.0, 0.0, 0.5, 0.5, 0.0\}$$
$$\theta = (0.5, 0.5)$$

LDA example 2-4

Suppose there are two sentences in *i*-th document. LDA starts with empty boxes to be filled in:

Credentials	***	

For the second cell,

- Draw a topic with probability θ_i. Suppose the second topic is selected
- Draw a word from $\mathcal{V}$ with probability ϕ_2. Suppose "maybe" is chosen

where

$$\mathcal{V} = \{\textit{credentials}, \textit{breach}, \textit{this}, \textit{maybe}, \textit{stolen}\}$$
$$\phi_1 = \{0.3, 0.4, 0.0, 0.0, 0.3\}$$
$$\phi_2 = \{0.0, 0.0, 0.5, 0.5, 0.0\}$$
$$\theta = (0.5, 0.5)$$

LDA example 2-5

Suppose there are two sentences in *i*-th document. LDA starts with empty boxes to be filled in ... and ...

Credentials	maybe	

Second position is filled with “maybe”

LDA example 2-5

Suppose there are two sentences in *i*-th document. LDA starts with empty boxes to be filled in ... and ...

Credentials	maybe	stolen

Third position is filled with “stolen”

LDA example 2-5

Suppose there are two sentences in *i*-th document. LDA starts with empty boxes to be filled in ... and ...

Credentials	maybe	stolen
this		

Fourth position is filled with “this”

LDA example 2-5

Suppose there are two sentences in *i*-th document. LDA starts with empty boxes to be filled in ... and ...

Credentials	maybe	stolen
this	maybe	

Fifth position is filled with “maybe”

LDA example 2-5

Suppose there are two sentences in *i*-th document. LDA starts with empty boxes to be filled in ... and ...

Credentials	maybe	stolen
this	maybe	breach

Last position is filled with “breach”

LDA in a typical application

The vocabulary size (V) is usually very large

The number of documents (M) is usually large

The number of topics (N) is relatively small

The word distribution ϕ is usually (approximately) sparse

$\Rightarrow$ LDA is a useful tool to summarize a large set of documents

Comment 1:
Theoretical justification for proposed metrics

Length score metric for i-th 10-K report is

$$LS_i = \sum_{k \in R_i} \sum_{j \in N} \phi_k^j w_k^i$$

where

- R_i: a set of words presented in i-th report and industry reports
- ϕ_k^j: a probability of k-th word from j-th topic word distribution
- w_k^i: the number of k-th word appeared in i-th 10-K report

Note that ϕ_k^j is from LDA estimation using the industry reports

If 10-K report contains more words that are frequently used in industry reports, then this metric gets higher number

It kind of makes sense, but can we justify it theoretically?

Theoretical justification

Consider the following conditional probability

$$p(i\text{-th 10-K document} \mid \text{industry reports})$$

- What is the probability of observing the *i*-th 10-K document conditional on observing industry reports
- How predictable *i*-th 10-K document given industry reports
- It gauges how close *i*-th 10-K document to industry reports
- This is known as “probability score”

We can compute this conditional probability using LDA model

Probability of seeing the word w (k-th word in R_i) in the i-th doc

$$\sum_{j \in N} p(word = w | topic = j) p(topic = j) = \sum_{j \in N} \phi_k^j \theta_j^i$$

where θ^i is a topic distribution for new document. Then, we have

$$p(i\text{-th 10-K document} \mid \text{industry reports}) = \prod_{k \in R_i} \left(\sum_{j \in N} \phi_k^j \theta_j^i \right)^{w_k^i}$$

"**Log probability score**" is then

$$\log p(i\text{-th 10-K document} \mid \text{industry reports}) = \sum_{k \in R_i} w_k^i \log \left(\sum_{j \in N} \phi_k^j \theta_j^i \right)$$

"**Log probability score**" is then

$$\log p(i\text{-th 10-K document} \mid \text{industry reports}) = \sum_{k \in R_i} w_k^i \log \left(\sum_{j \in N} \phi_k^j \theta_j^i \right)$$

A crude approximation to it leads us to ...

$$\begin{aligned} \sum_{k \in R_i} w_k^i \log \left(\sum_{j \in N} \phi_k^j \theta_j^i \right) &\approx \sum_{k \in R_i} w_k^i \left(\sum_{j \in N} \phi_k^j \theta_j^i - 1 \right) \\ &= \sum_{k \in R_i} w_k^i \sum_{j \in N} \phi_k^j \theta_j^i - \sum_{k \in R_i} w_k^i \\ &= \sum_{k \in R_i} \sum_{j \in N} \phi_k^j w_k^i \theta_j^i - W_i \end{aligned}$$

Log probability score is

$$\text{Log probability score}_i \approx \sum_{k \in R_i} \sum_{j \in N} \phi_k^j w_k^i \theta_j^i - W_i$$

Original LS is

$$\text{Length score}_i = \sum_{k \in R_i} \sum_{j \in N} \phi_k^j w_k^i$$

Remarks

1. If θ_j^i is ignored, then it is $LS_i - W_i$.
2. Log score has a built-in penalization for the longer document. But, the first term increases as W_i increases as well.
3. We can divide it by W_i, which results in average predictive score, and it becomes MS_i,

$$\text{Average log probability score}_i \approx \underbrace{\sum_{k \in R_i} \sum_{j \in N} \phi_k^j w_k^i \theta_j^i / W_i}_{=MS_i \text{ if } \theta \text{ is ignored}} - 1$$

Bonus: Full-Bayes characterization

Predictive score with given ϕ and θ:

$$p(i\text{-th 10-K document} \mid \text{industry reports}) = \prod_{k \in R_i} \left(\sum_{j \in N} \phi_k^j \theta_j^i \right)^{w_k^i}$$

Bayes can handle parameter uncertainty in a simple manner:

- Replace $\sum_{j \in N} \phi_k^j \theta_j^i$ with the following integration

$$\int_\phi \int_\theta \left(\sum_{j \in N} \phi_k^j \theta_j \right) \underbrace{p(\phi, \theta | \text{industry reports})}_{\text{what LDA estimates}} d\phi d\theta$$

- This quantity can be approximated by the Monte Carlo integration

Comment 2:
Validation for internal consistency

The author performs an external validation by comparing proposed metrics to ones based on “cosine” similarity

A more direct way to validate internal consistency:

- Recall that three metrics are only computed for hand-collected paragraphs that are relevant for cybersecurity
- Suppose you compute three metrics for other paragraphs from the same report
- Scores should be higher for hand-collected paragraphs than scores based on other paragraphs from the same report

This is particularly important because the author uses “all topics” obtained from industry reports

- There could be other topics that are not about cybersecurity
- It may be better to start with a large number of topics, then use only those that are related to cybersecurity when computing metrics

Conclusion

It is an interesting and novel application of LDA to an important issue

My comments are ...

- Proposed scores have a close connection to “log-predictive score”
- Derivation reveals that one may need to take into account topic distribution (θ)
- A more direct validation can be performed