Prepared by TEAM3

Everything of Handong

from 1995 to 2024 (Analysis of Handong News)

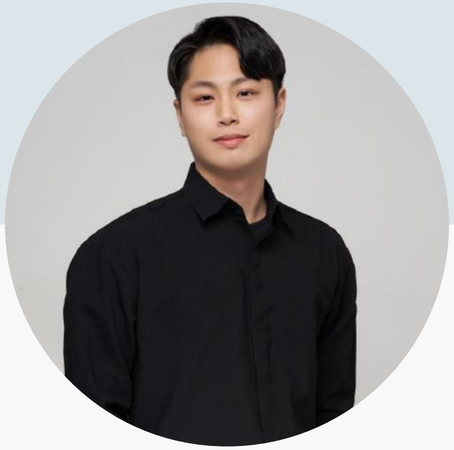
3 December, 2024

Team03 21900759 Choe Jaeseong, 21900844 Yugay Dmitriy, 22100727 Cheon JeongWon

Team Members

Dmitriy Yugay

https://github.com/yudm3



Jaeseong Choe

https://github.com/sorrychoe

JeongWon Cheon

https://github.com/garden1000

Introduction

"What do you think is the most frequently discussed topic in news about Handong University?"

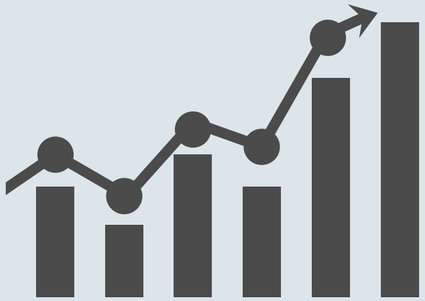
Research Objectives

To find out how the world views Handong University through reports related to Handong University that have been published since its establishment.

Research Questions

- Q1. What words did the media usually use to report on Handong University?
- Q2. Is the image of Handong University described by the media positive or negative?
- Q3. What are the main topics that emerged from reports related to Handong University?
- Q4. Does the reporting on Handong University have any influence on the media partisanship?

Research Method



Frequency Analysis

- Analysis of the relative frequency of the text
- TF-IDF was used for weights.

Sentiment Analysis

- Lexicon-based sentiment analysis
- A KNU emotional dictionary was used for this analysis.

Topic Modeling

- Statistical methodology for identifying potential topics in a document
- In this analysis, a structural topical model (STM) was used.

TF-IDF

$$\mathbf{tf}(t, d) = \frac{f_d(t)}{\max_{w \in d} f_d(w)}$$

$$\mathbf{idf}(t, D) = \ln\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right)$$

$$\mathbf{tfidf}(t, d, D) = \mathbf{tf}(t, d) \cdot \mathbf{idf}(t, D)$$

$$\mathbf{tfidf}'(t, d, D) = \frac{\mathbf{idf}(t, D)}{|D|} + \mathbf{tfidf}(t, d, D)$$

$f_d(t) :=$ frequency of term t in document d

$D :=$ corpus of documents

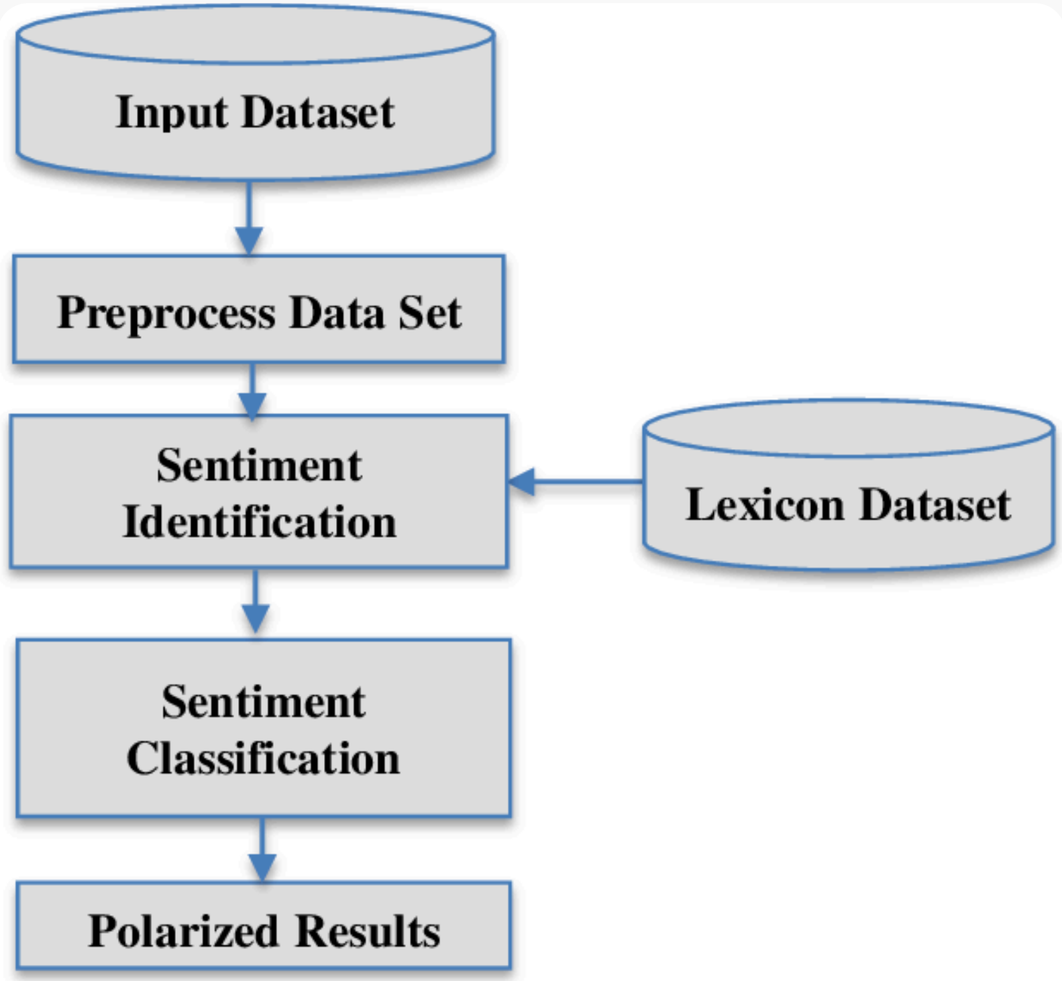
What is the TF-IDF?

- TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical method used to evaluate the importance of a word in a document relative to a collection of documents (corpus).
- It balances the frequency of a term in a specific document with its rarity across the entire corpus.

Representation of TF-IDF

- TF (Term Frequency): The frequency at which a specific word appears in a particular document.
- IDF (Inverse Document Frequency): Represents the rarity of a word across the entire document set; less frequent words have higher values.
- TF-IDF Calculation: Measures the importance of a word by multiplying TF and IDF.

Sentiment Analysis



What is the Sentiment Analysis

- Sentiment Analysis is a technique used to detect and interpret emotions expressed within textual data, categorizing sentiments as positive, negative, or neutral.
- Sentiment analysis is broadly divided into lexicon-based analysis and machine learning-based analysis. In this analysis, we conducted a lexicon-based approach.

About Lexicon-Based Sentiment Analysis

- The lexicon-based analysis derives sentiment scores by attaching a sentiment dictionary to tokenized words. The derived sentiment scores are averaged per document.
- In this analysis, we utilized the KNU Korean Sentiment Lexicon developed by the Data Intelligence Lab at Kunsan National University.

KNU Sentiment Lexicon

ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org

J Intell Inform Syst 2018 December: 24(4): 219~240 http://dx.doi.org/10.13088/jiis.2018.24.4.219

Bi-LSTM 기반의 한국어 감성사전 구축 방안*

박상민
군산대학교 소프트웨어융합공학과
(b1162@kunsan.ac.kr)

나철원
군산대학교 소프트웨어융합공학과
(ncw0034@kunsan.ac.kr)

최민성
군산대학교 소프트웨어융합공학과
(alstjd517@kunsan.ac.kr)

이다희
군산대학교 소프트웨어융합공학과
(dahee@kunsan.ac.kr)

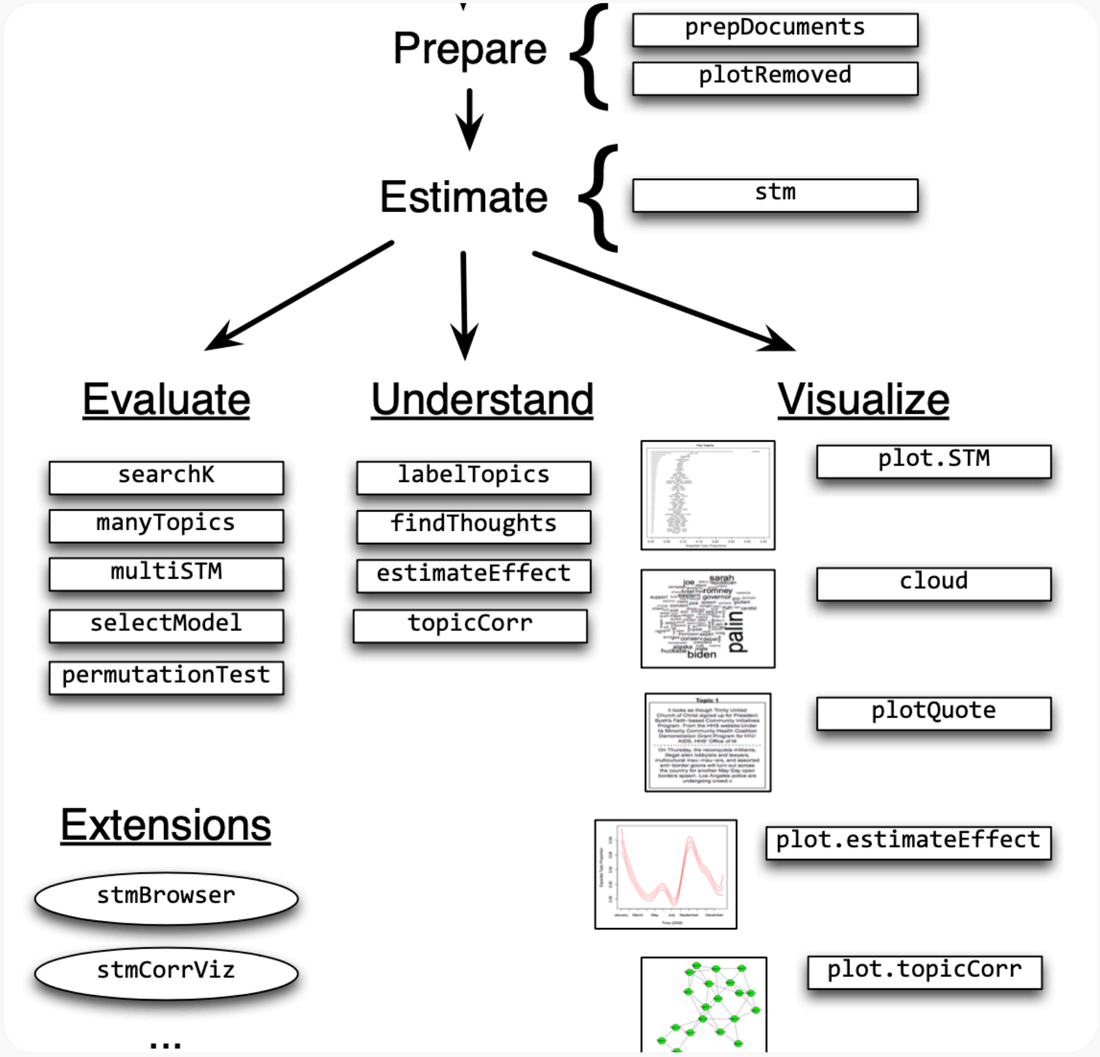
온병원
군산대학교 소프트웨어융합공학과
(bwon@kunsan.ac.kr)

감성사전은 감성 어휘에 대한 사전으로 감성 분석(Sentiment Analysis)을 위한 기초 자료로 활용된다. 이와 같은 감성사전을 구성하는 감성 어휘는 특정 도메인에 따라 감성의 종류나 정도가 달라질 수 있다. 예를 들면, '슬프다'라는 감성 어휘는 일반적으로 부정의 의미를 나타내지만 영화 도메인에 적용되었을 경우 부정의 의미를 나타내지 않는다. 그렇기 때문에 정확한 감성 분석을 수행하기 위해서는 특정 도메인에 알맞은 감성사전을 구축하는 것이 중요하다. 최근 특정 도메인에 알맞은 감성사전을 구축하기 위해 범용 감성 사전인 오픈한글, SentiWordNet 등을 활용한 연구가 진행되어 왔으나 오픈한글은 현재 서비스가 종료되어 활용이 불가능하며, SentiWordNet은 번역 간에 한국 감성 어휘들의 특징이 잘 반영되지 않는다는 문제점으로 인해 특정 도메인의 감성사전 구축을 위한 기초 자료로써 제약이 존재한다. 이 논문에서는 기존의 범용 감성사전의 문제점을 해결하기 위해 한국어 기반의 새로운 범용 감성사전을 구축하고 이를 KNU 한국어 감성사전이라 명명한다. KNU 한국어 감성사전은 표준국어대사전의 뜻풀이의 감성을 Bi-LSTM을 활용하여 89.45%의 정확도로 분류하였으며 긍정으로 분류된 뜻풀이에서는 긍정에 대한 감성 어휘를, 부정으로 분류된 뜻풀이에서는 부정에 대한 감성 어휘를 1-gram, 2-gram, 어구 그리고 문형 등 다양한 형태로 추출한다. 또한 다양한 외부 소스(SentiWordNet, SenticNet, 감정동사, 감성사전0603)를 활용하여 감성 어휘를 확장하였으며 온라인 텍스트 데이터에서 사용되는 신조어, 이모티콘에 대한 감성 어휘도 포함하고 있다. 이 논문에서 구축한 KNU 한국어 감성사전은 특정 도메인에 영향을 받지 않는 14,843개의 감성 어휘로 구성되어 있으며 특정 도메인에 대한 감성사전을 효율적이고 빠르게 구축하기 위한 기초 자료로 활용될 수 있다. 또한 딥러닝의 성능을 높이기 위한 입력 자질로써 활용될 수 있으며, 기본적인 감성 분석의 수행이나 기계 학습을 위한 대량의 학습 데이터 세트를 빠르게 구축에 활용될 수 있다.

What is the KNU Sentiment Lexicon

- One of the most popular Korean sentiment dictionary.
- The KNU Korean Sentiment Lexicon contains a total of 14,843 expressions, including words from the Standard Korean Dictionary, idiomatic expressions, and abbreviations.
- It includes polarity and intensity values of positive, negative, and neutral for each vocabulary word.
- Sentiment scores range from -2 to 2; the more positive the sentiment, the closer the score is to 2, and the more negative, the closer it is to -2.

STM



What is the STM?

- One of the Topic Models used in social science research.
- Structural Topic Model(STM) incorporates document-level metadata into the topic modeling process.
- It allows the inclusion of external information (metadata) to influence the document's topic distribution.

Why STM Was Introduced

- Before STM, Topic models like LDA treated documents independently and ignored document-specific covariates.
- STM was introduced from the paper "A model of text for experimentation in the social sciences." of Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016).
- STM allows incorporating document-level metadata (e.g., author, date) to influence topic discovery.

Process of STM

1. **Covariate Effect on Topic Prevalence**:

   - For each document d, the topic distribution θ_d is influenced by covariates X_d. The topic proportions θ_d are drawn from a logistic normal distribution:

     $$\theta_d \sim \text{LogisticNormal}(\mu_d, \Sigma)$$

     where $\mu_d = X_d\gamma$, and Σ is the covariance matrix.

2. **Topic Assignment**:

   - For each word w_{dn} in document d, draw a topic assignment z_{dn} from a multinomial distribution:

     $$z_{dn} \sim \text{Multinomial}(\theta_d)$$

3. **Covariate Effect on Topic Content**:

   - The word distribution β_k for topic k depends on the covariates Y_d. The distribution over words is modeled as a multinomial logistic regression:

     $$\beta_{d,k,v} \propto \exp(m_v + \kappa_k^{(t)} + \kappa_v^{(c)} + \kappa_{kv}^{(i)})$$

     where:

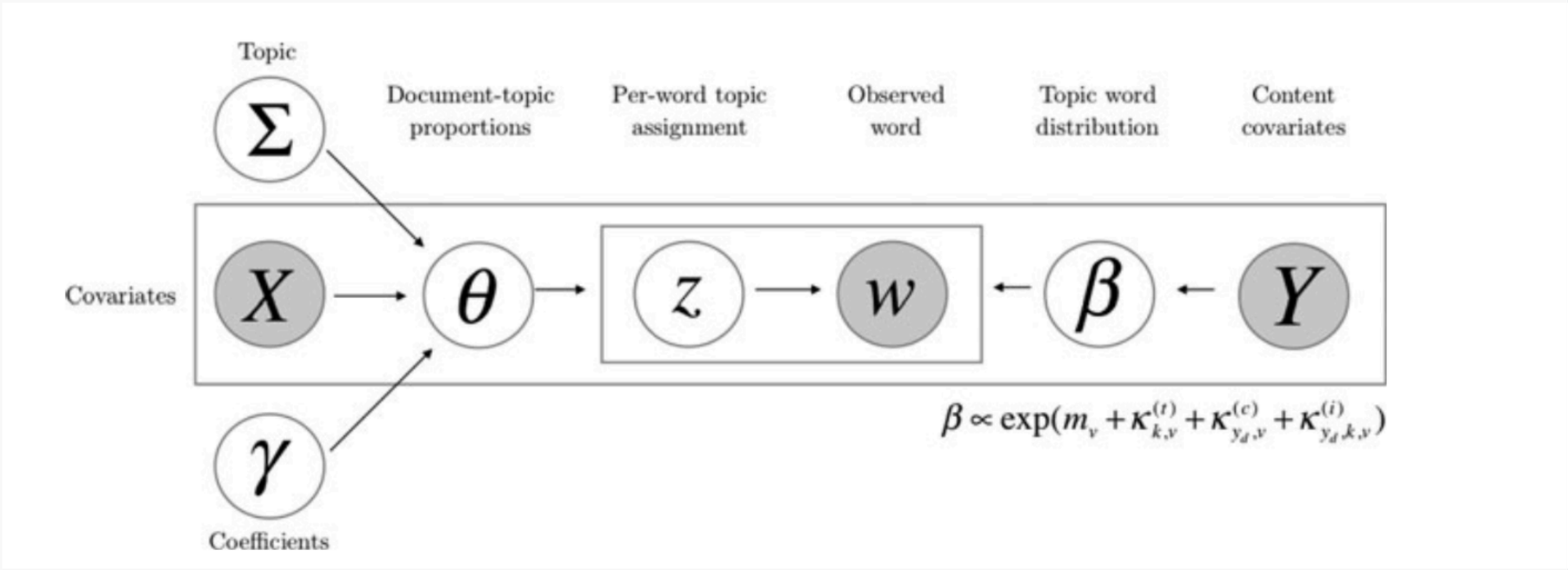
   - m_v is the marginal log frequency of term v,
   - $\kappa_k^{(t)}$ is the topic-specific effect,
   - $\kappa_v^{(c)}$ is the covariate-specific effect,
   - $\kappa_{kv}^{(i)}$ is the topic-covariate interaction effect.

4. **Word Generation**:

   - For each word w_{dn}, draw a word from the multinomial distribution based on the assigned topic z_{dn}:

     $$w_{dn} \sim \text{Multinomial}(\beta_{z_{dn}})$$

Process of STM



Mathematical Representation of STM

Mathematical Representation

- **Topic Prevalence Model**: The topic proportions θ_d are modeled as a logistic normal distribution:

$$\theta_d = \frac{\exp(\eta_d)}{\sum_k \exp(\eta_d, k)}$$

  where $\eta_d \sim \mathcal{N}(X_d\gamma, \Sigma)$.

- **Topic Content Model**: The word distribution for each topic β_k is influenced by the covariates Y_d:

$$\beta_{k,v} \propto \exp(m_v + \kappa_k^{(t)} + \kappa_v^{(c)} + \kappa_{kv}^{(i)})$$

Inference

STM uses **variational expectation-maximization (EM)** for approximate inference due to the non-conjugacy of the logistic normal distribution with the multinomial likelihood.

- **Variational E-Step**: The variational posterior for η_d is approximated using a Laplace approximation:

$$q(\eta_d) \approx \mathcal{N}(\lambda_d, \nu_d)$$

  where λ_d is the mode of η_d, and ν_d is the Hessian of the log-posterior.

- **Variational M-Step**: The M-step maximizes the ELBO with respect to the model parameters γ, κ, Σ. The updates for γ are obtained through linear regression, while β is updated using a multinomial logistic regression.

Analytics Process

Data Collection and Preprocessing

- Extracted news articles related to Handong University from BigKinds spanning from 1995 to October 2024.
- A total of 7,857 articles were collected.
- The texts were already tokenized using BigKinds' own tokenizer, "Bareun".

Frequency Analysis

- Conducted frequency analysis based on TF-IDF.
- Calculated TF-IDF scores using the tidytext library (bind_tf_idf() function).
- Presented the analysis results in a table.

Sentiment Analysis

- Calculated sentiment scores by linking the KNU Sentiment Lexicon at the word level.
- Averaged sentiment scores per document to minimize the influence between document length and sentiment scores.
- Visualized sentiment scores using a time series chart showing annual averages.

Analytics Process

Topic Modeling

- Extracted topics using STM (Structural Topic Modeling).
- Treated the university president's tenure and the media partisanship as covariates, and conducted an ANCOVA (Analysis of Covariance) between topics and covariates.

Visualization

- Performed visualization of the topic modeling results.

R1. Frequency Analysis Result

=> Perform “TF-IDF” - based weighted frequency analysis

Key Keywords:

Diplomacy and International Politics: North Korea, United States, Trump

Religious Context: Pastor, Church, God

Pohang and Local Community: Pohang, Earthquake, Support, Region

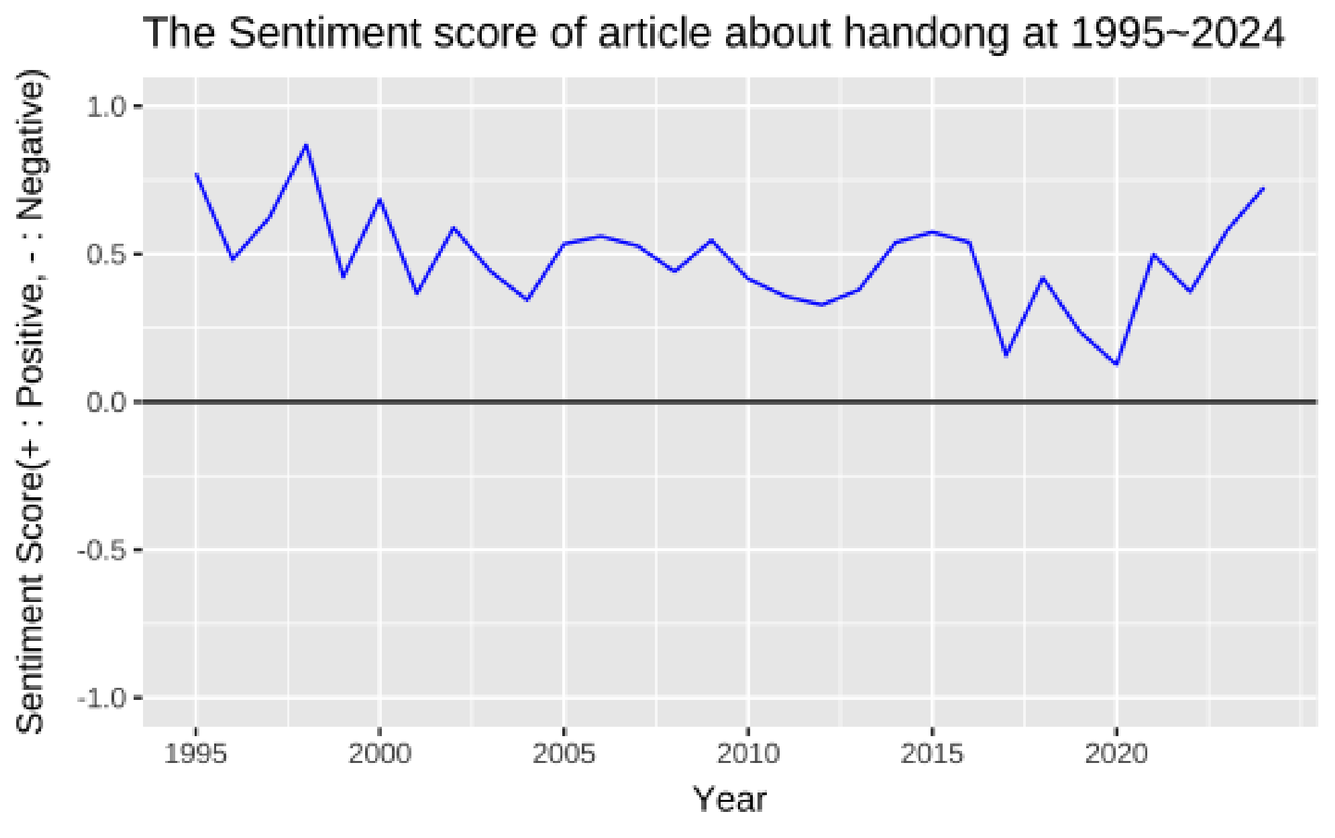
Interpretation:

The findings suggest that Handong University is prominently linked to international, religious, and regional contexts, reflecting its multifaceted image and the key narratives surrounding it.

```
# A tibble: 20 × 2
   words  score
   <chr>  <dbl>
 1 북한    72.8
 2 대학    67.2
 3 미국    51.0
 4 총장    45.2
 5 목사    39.4
 6 대통령  38.2
 7 교육    36.9
 8 포항    36.2
 9 한국    35.7
10 교회    34.5
11 하나님  31.9
12 중국    31.9
13 정부    30.9
14 학생    30.7
15 교수    30.2
16 트럼프  29.7
17 지진    29.5
18 지원    29.4
19 협상    28.2
20 지역    28.1
```

R2. Sentiment Analysis Result

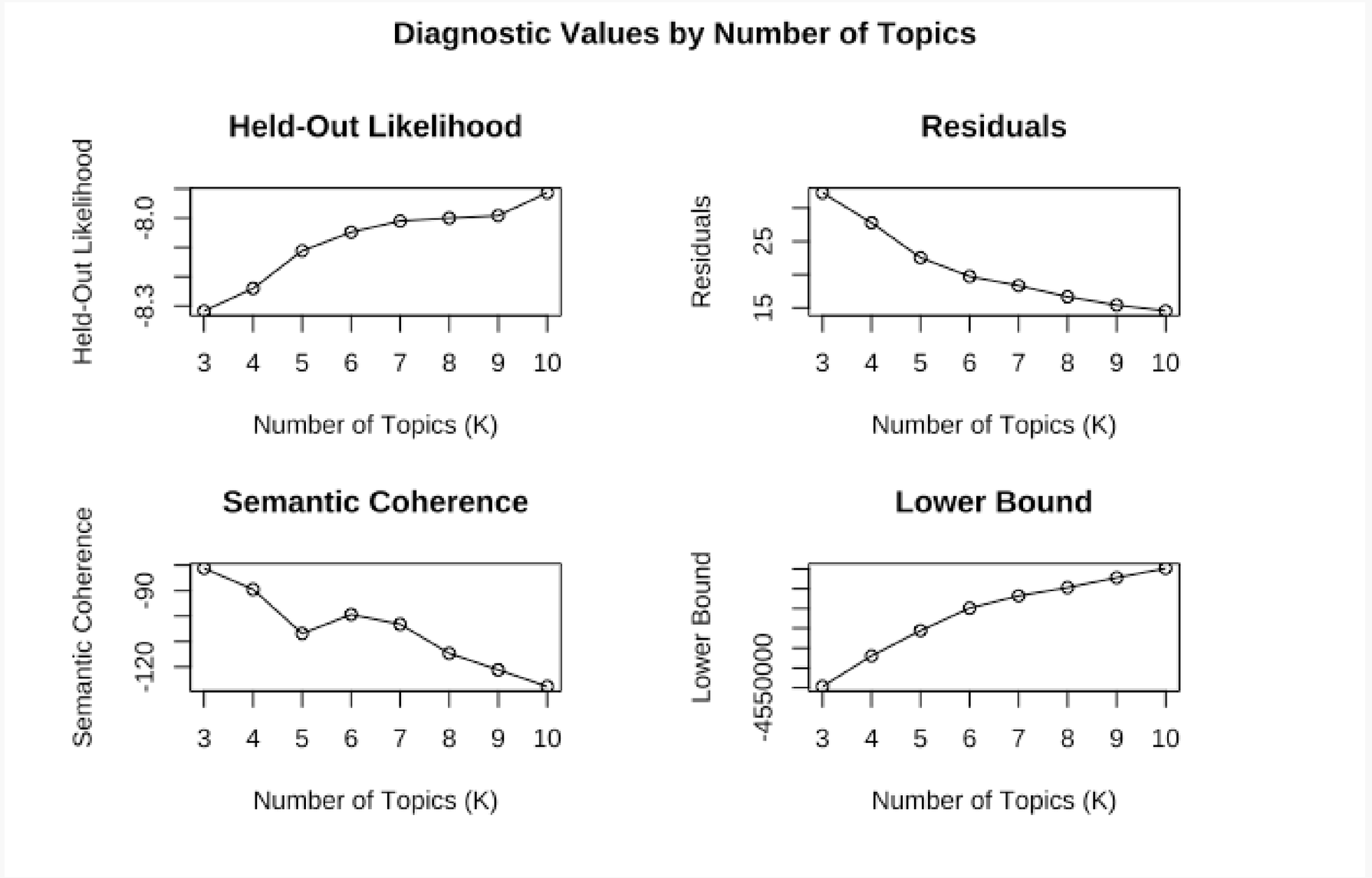
=> Conduct lexicon-based sentiment analysis.



Interpretation

Tone of news articles related to Handong University generally carried a POSITIVE sentiment.

R3. Topic Modeling Result



R3. Topic Modeling Result

Methods to Determine K:

- **STM: Use Held-out Likelihood, Semantic Coherence, Lower Bound, and Residual.**

Key Metrics:

- **Higher Held-out Likelihood, Semantic Coherence, and Lower Bound and lower Residual**

Findings:

- **Held-out Likelihood, Lower Bound, and Residual showed consistent trends.**
- **Semantic Coherence declined, suggesting a shared theme (Handong University).**

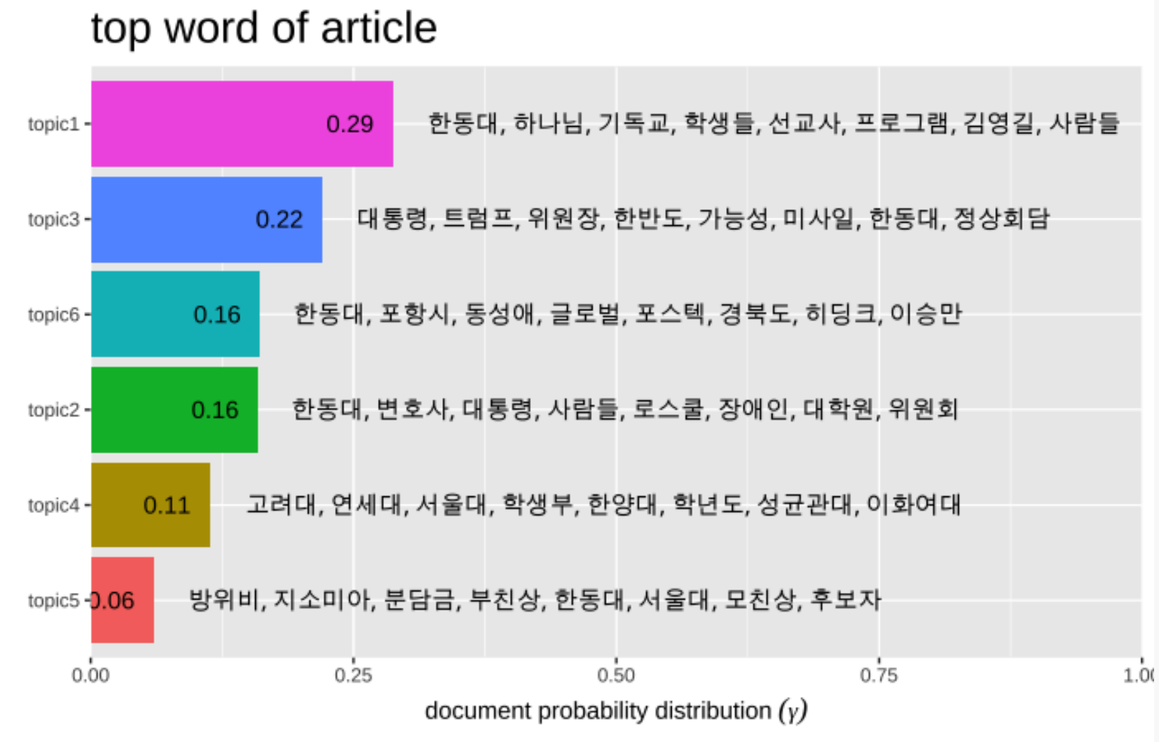
Conclusion:

- **Six topics**

R3. Topic Modeling Result

Topic Word Analysis:

Keywords help define the main themes present in the articles



Topic 1. Christian Spirit of Handong

Topic 2. Legal Issues Related to Handong

Topic 3. Political and Diplomatic Briefings by Professors

Topic 4. University Admissions

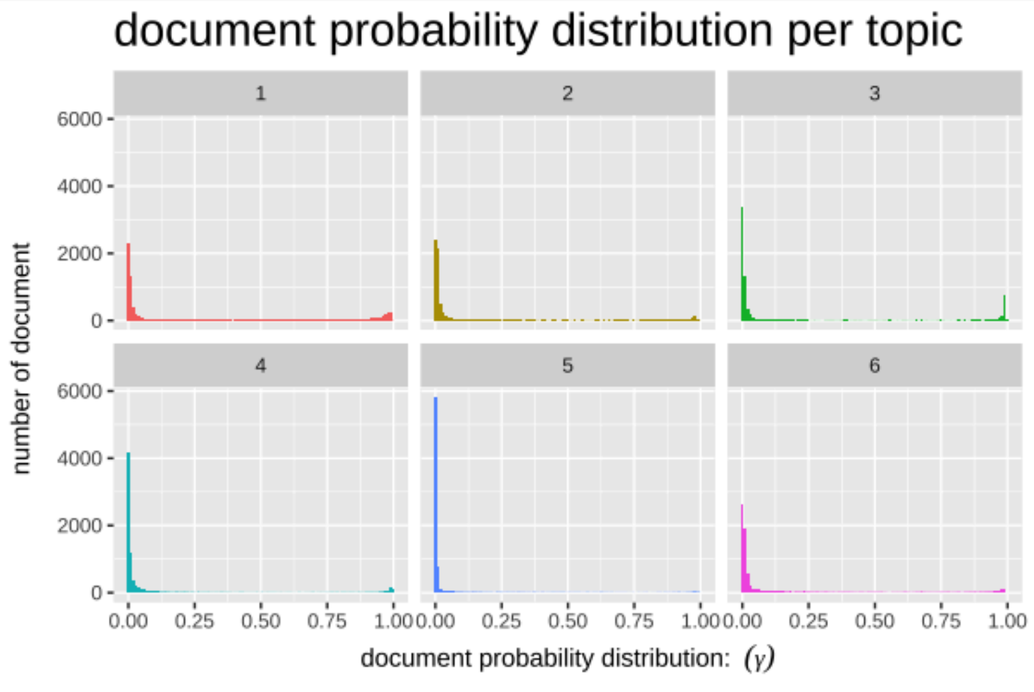
Topic 5. News About Handong Members

Topic 6. External Issues of Handong University

R3. Topic Modeling Result

Gamma distribution (γ):

Represents the probability of each document belonging to a particular topic.



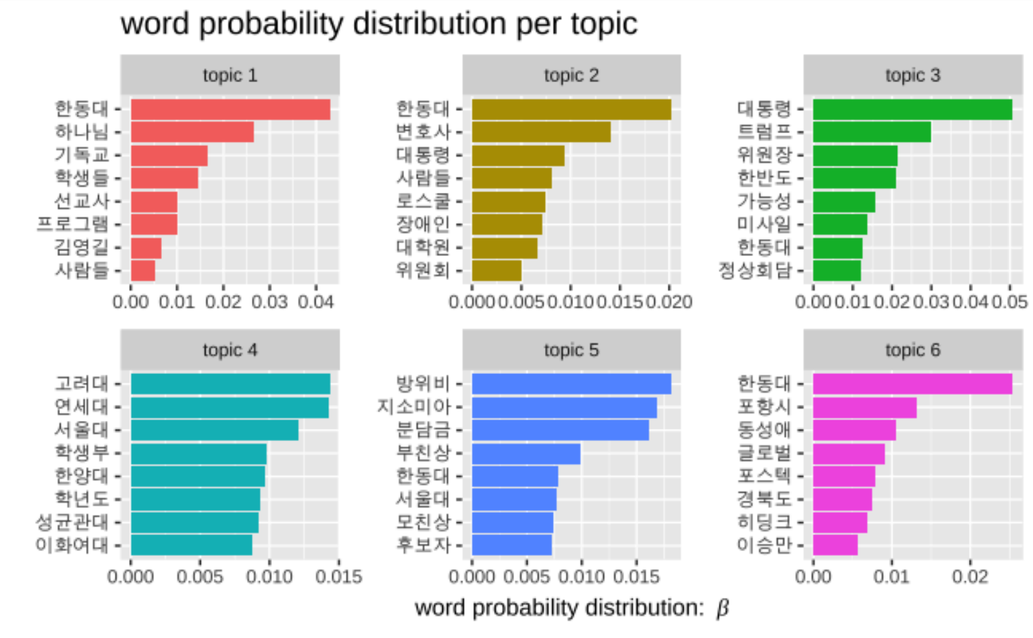
Key Insights:

- **Most documents exhibit low probabilities for each topic, indicating that topics are highly specific.**
- **A small number of documents have probabilities near 1.0, showing they are strongly associated with a single topic.**
- **This demonstrates that the STM model effectively identifies distinct themes in the data.**

R3. Topic Modeling Result

Beta Distribution (β):

Represents the probability of specific words occurring in each topic.



Key Insights:

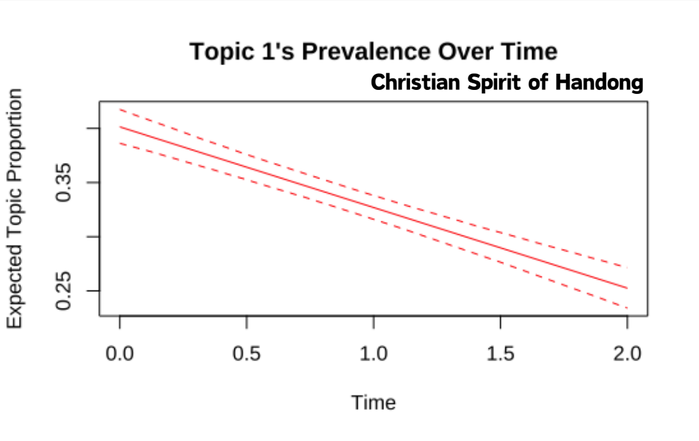
- **Each topic is characterized by a unique set of high-probability words, reflecting distinct themes.**
- **The top words for each topic align with its primary focus, such as Christianity, legal issues, or university admissions.**
- **This demonstrates the STM model's capability to uncover meaningful patterns in word associations across topics.**

R3. Topic Modeling Result

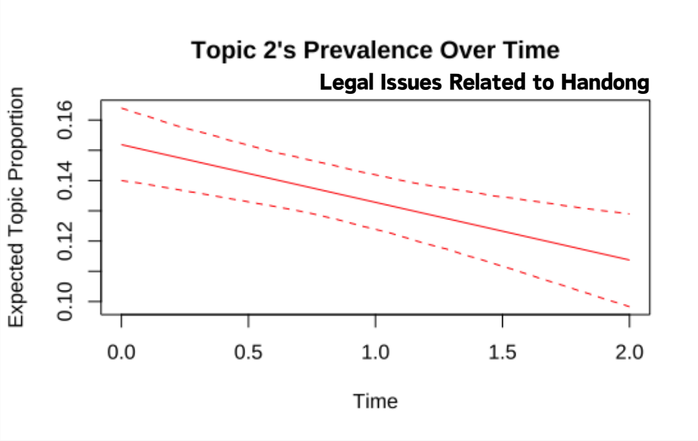
Topic Prevalence over time:

Observe how the prevalence of topics has changed over time.

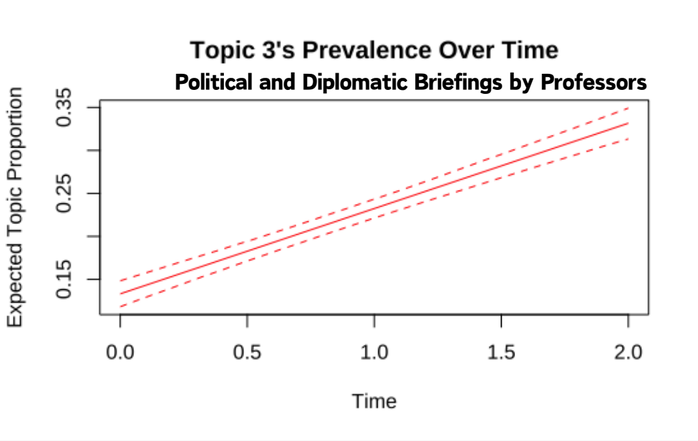
Tenures of presidents Kim Young-Gil(0), Jang Soon-Heung(1), and Choi Do-Sung(2)



- **Decreasing trend**
- **More prominent in earlier years but appear less frequently in recent media**



- **Moderate decline**
- **Decrease is less pronounced compared to Topic 1**



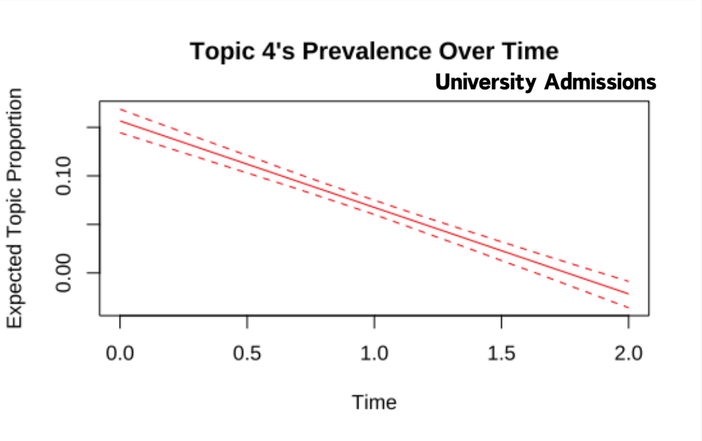
- **Significant Increase**
- **Active engagement in media discussions on political and diplomatic topics**

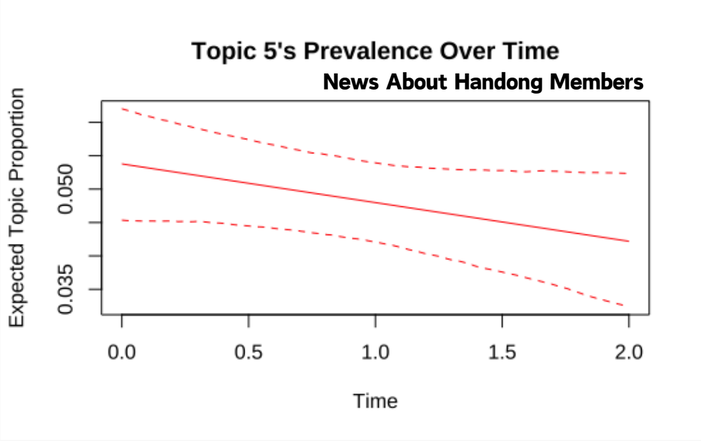
R3. Topic Modeling Result

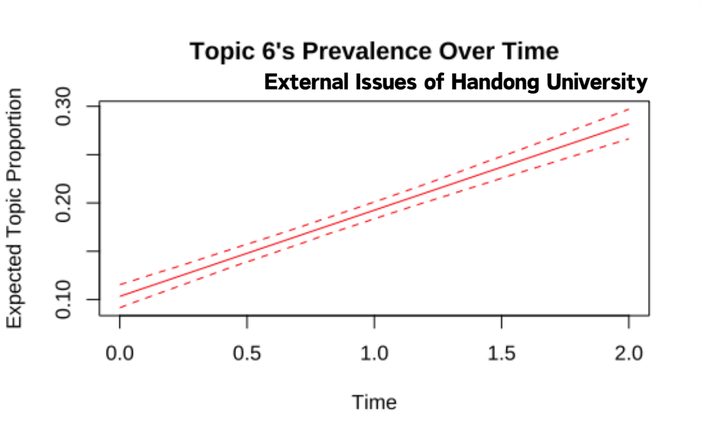
Topic Prevalence over time:

Observe how the prevalence of topics has changed over time.

Tenures of presidents Kim Young-Gil(0), Jang Soon-Heung(1), and Choi Do-Sung(2)





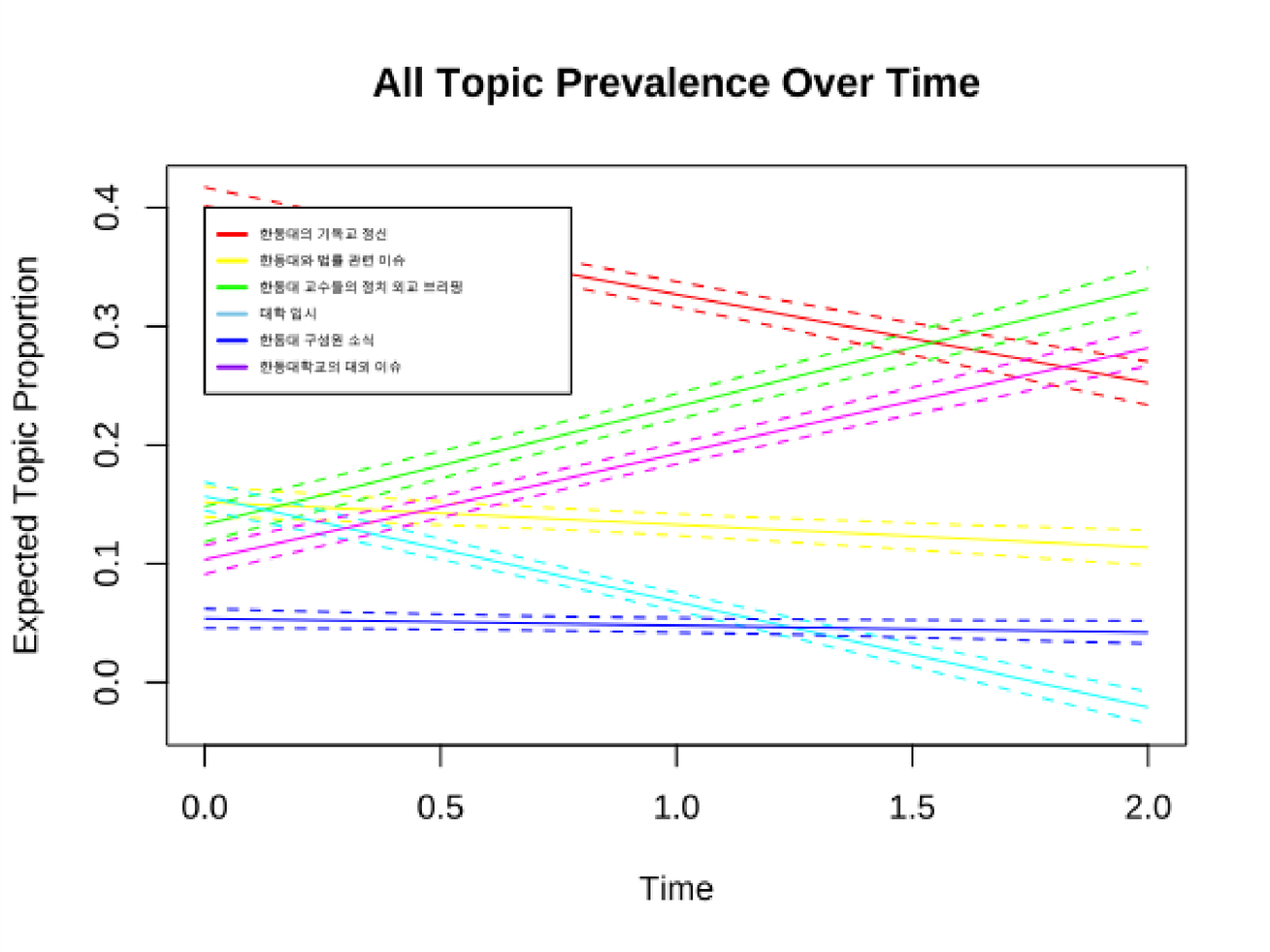


- **Prominent in earlier years but has almost vanished recently**
- **Focus away from university admissions**

- **Slow decline over time**
- **Despite the decline, remains a consistently reported topic**

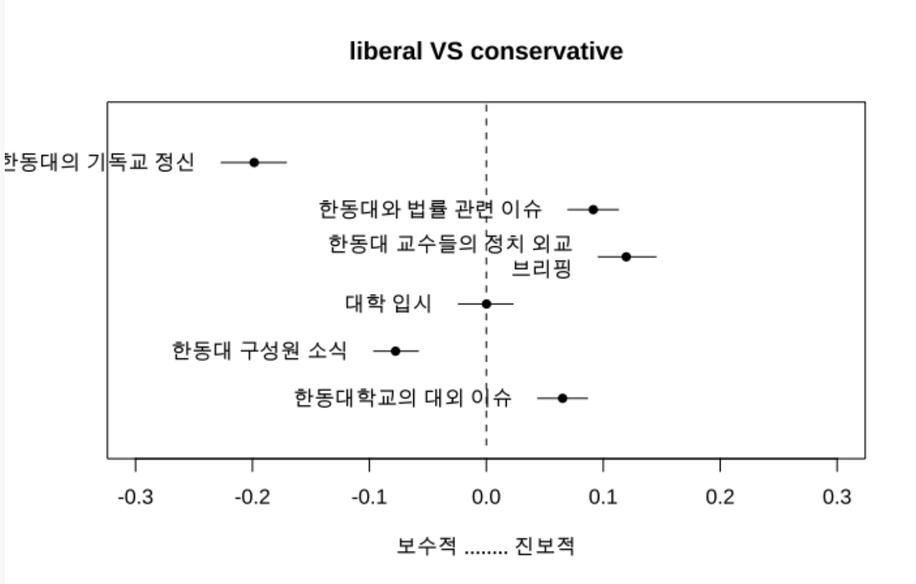
- **Sharply increase**
- **External achievements appear more frequently in the media**

R3. Topic Modeling Result



R3. Topic Modeling Result

Political Orientation of Each Topic



Conservative: Chosun, JoongAng, DongA

Progressive: Kyunghyang, Hankyoreh

*widely accepted classifications in media studies

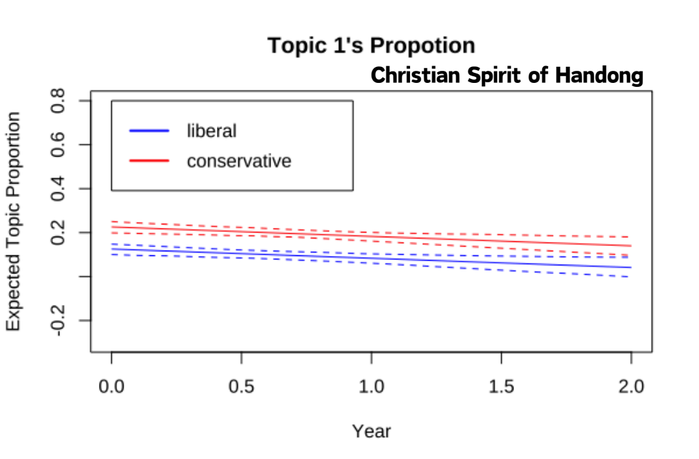
- **Conservative-leaning Topics:**
  - -Christian Spirit: Strongly highlighted in conservative media.
  - -Handong Members: Minimal bias across media.
- **Liberal-leaning Topics:**
  - -Political and Diplomatic Brefings: Frequently covered in liberal media.
  - -External Achievements: Slightly leaned toward liberal media.
  - -Legal Issues: Slightly leaned toward liberal media.
- **Neutral Topics:**
  - -University Admissions: Evenly covered by both media types.

R3. Topic Modeling Result

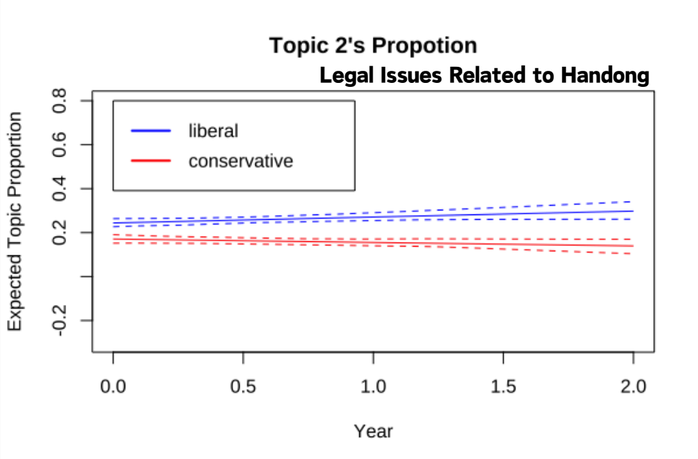
Topic Proportion over time

Observe how the proportion of topics has changed over time and by the political stance of the media

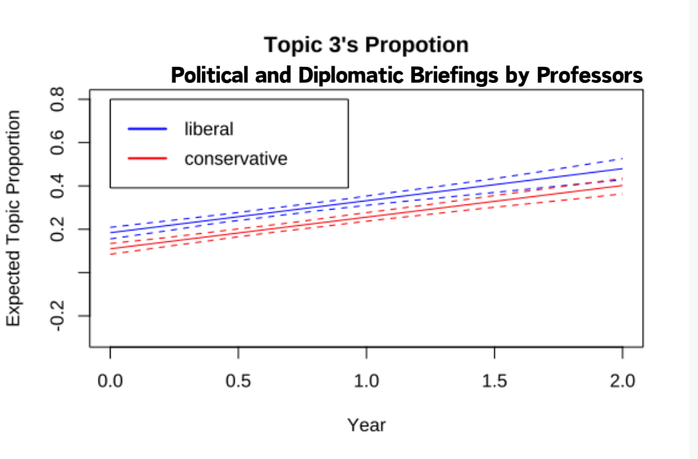
Tenures of presidents Kim Young-Gil(0), Jang Soon-Heung(1), and Choi Do-Sung(2)



- **Decreasing Trend**
- **Conservative media consistently reports more on this topic than liberal media**



- **Relatively stable**
- **Liberal media consistently shows higher coverage than conservative media**



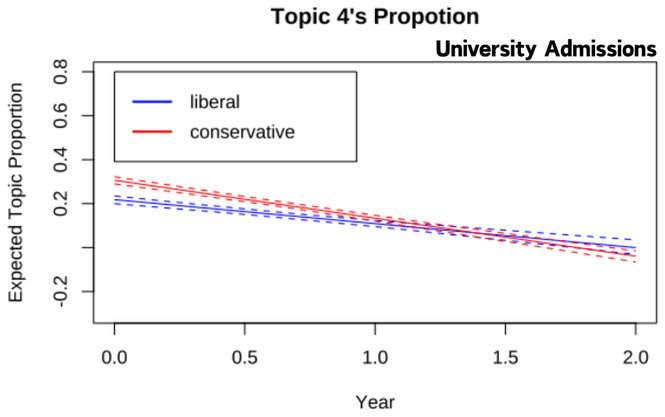
- **Significant Increase**
- **Liberal media shows a higher proportion due to Handong professors' contributions.**

R3. Topic Modeling Result

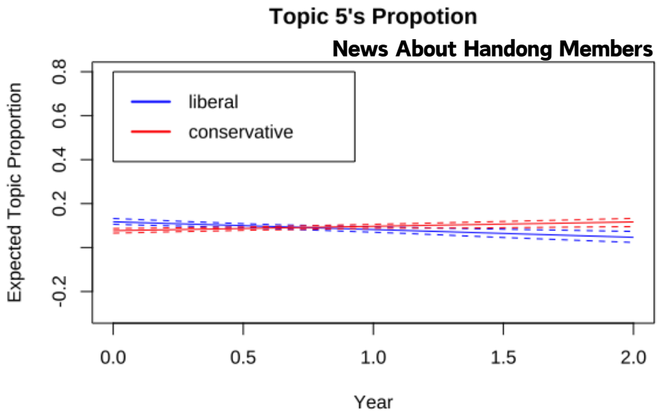
Topic Proportion over time

Observe how the proportion of topics has changed over time and by the political stance of the media

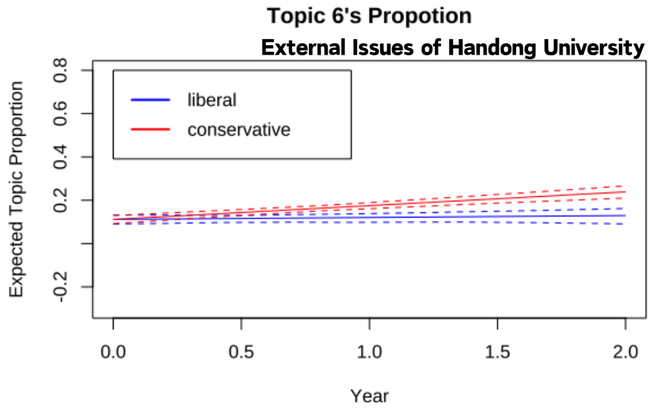
Tenures of presidents Kim Young-Gil(0), Jang Soon-Heung(1), and Choi Do-Sung(2)



- **Decreasing Trend**
- **Liberal and conservative media exhibit similar proportions for this topic**



- **Stable Trend**
- **Both liberal and conservative media cover this topic similarly over time**



- **Increasing Trend**
- **Conservative media shows a marginally higher proportion compared to liberal media**

Answer to Research Questions

- Q1. What words did the media usually use to report on Handong University?

  => The keyword is about International Politics, Religion & Local Community

- Q2. Is the image of Handong University described by the media positive or negative?

  => It was generally positive.

- Q3. What are the main topics that emerged from reports related to Handong University?

  => The main topic is 6 which we previously mentioned.

- Q4. Does the reporting on Handong University have any influence on media partisanship?

  => Usually Yes, In particular, Handong's Christianity was mainly covered by conservative media.

Thank you