

# Addressing Controversial Topics in Search Engines

---

The Oral Exam of  
**Yamen Ajjour**

To Obtain the Academic Degree of  
**Dr. rer. nat.**

# Addressing Controversial Topics in Search Engines

- Motivation
- Identifying argumentative questions
- Topic bias in argument corpora
- Identifying argument frames
- Conclusion

# Motivation

Use cases where good arguments are needed.



Student



Lawyer




Politician



Marketing Company

# Motivation

Search engines are good at answering factual questions.



The image shows a Google search interface. The search bar contains the text "where was marijuana first legalized". Below the search bar, there are navigation options: "Alle", "News", "Bilder", "Videos", "Shopping", and "Mehr". The search results show "Ungefähr 58.400.000 Ergebnisse (0,71 Sekunden)". The first result is titled "California" and the text below it states: "California was the first state to legalize medical marijuana in 1996. Since then, the medical use of cannabis has been legalized in 40 states and the District of Columbia. The recreational or adult-use of cannabis has been approved in DC and 23 states."

Google

where was marijuana first legalized

Alle News Bilder Videos Shopping Mehr Suchfilter

Ungefähr 58.400.000 Ergebnisse (0,71 Sekunden)

## California

California was the first state to legalize medical marijuana in 1996. Since then, the medical use of cannabis has been legalized in 40 states and the District of Columbia. The recreational or adult-use of cannabis has been approved in DC and 23 states.

# Motivation

Search engines struggle at delivering all perspectives on controversial topics.

The screenshot shows a Google search interface with the query "should marijuana remain illegal". The search results are filtered to show only "Alle" (All) results. The top results are:

- drugrehab.us** (https://www.drugrehab.us > news) - **PRO**  
**Why Recreational Marijuana Should Stay Illegal - Drug Rehab**  
Many people think marijuana **isn't harmful enough** to be illegal, but this is dangerous - See the reasons why recreational marijuana should stay illegal now!
- National Institutes of Health (.gov)** (https://www.ncbi.nlm.nih.gov > P...) - **PRO**  
**More Reasons States Should Not Legalize Marijuana**  
von ST Wilkinson · 2013 · Zitiert von: 16 — Medical **marijuana should be** subject to the same rigorous approval process as other medications prescribed by physicians. Legalizing recreatio...  
Recreational Marijuana · Myth: Marijuana is Not Addictive · Effects on Cognition
- Liberty University** (https://www.liberty.edu > 2022/03) - **PRO**  
**Cons of Legalizing Recreational Marijuana Use**  
07.03.2022 — Despite the growing acceptance, decriminalizing marijuana should be a federal issue, and **it should not be legal to use recreationally.**
- Change.org** (https://www.change.org > guam-g...) - **PRO**  
**Petition · Marijuana should remain illegal!**  
Marijuana should remain illegal. **The federal government is planning to legalize marijuana,** however, it's policy on marijuana is like an octopus.

# Motivation

Argument retrieval systems retrieve pro and con arguments for a query.

**args** should marijuana remain illegal 🔍

All Discussions News People Pro vs. con view ▾ 1513 arguments retrieved in 1.0 ms

**PRO**

[This debate will be about whether weed should be...](#)  
► Show full argument  
This debate will be about whether weed **should** be legalized or **remain illegal**. ... For round 1, when accepting my challenge, Introduce yourself, but the arguing will not begin until the second round.  
<https://www.debate.org/debates/Marijuana-should-remain-illegal/1/> score ▾

[Well, there isn't anything that especially demonizes...](#)  
► Show full argument  
Well, there isn't anything that especially demonizes **marijuana**, but what lifting the ban of **marijuana** would do is 1: Increase use, whether it be recreational or medicinal, Leaving millions susceptible to falling prey to ...  
<https://www.debate.org/debates/Marijuana-should-remain-illegal/1/> score ▾

[This "gift of nature" is a cause of several thousand...](#)  
► Show full argument  
This "gift of nature" is a cause of several thousand employment terminations, incarcerations, relationship breakups, car accidents, and more. ... The fact that it "grows naturally" doesn't excuse the fact it is an abused drug, and ...  
<https://www.debate.org/debates/Marijuana-should-remain-illegal/1/> score ▾

**CON**

[I am for the legalization of recreational and medical...](#)  
► Show full argument  
I am for the legalization of recreational and medical uses of **Marijuana**. ... **Marijuana-Cannabis illegal**- In this context, to have the sale, possession, distribution and growth **illegal** to be punished by the law  
<https://www.debate.org/debates/Marijuana-should-remain-illegal/1/> score ▾

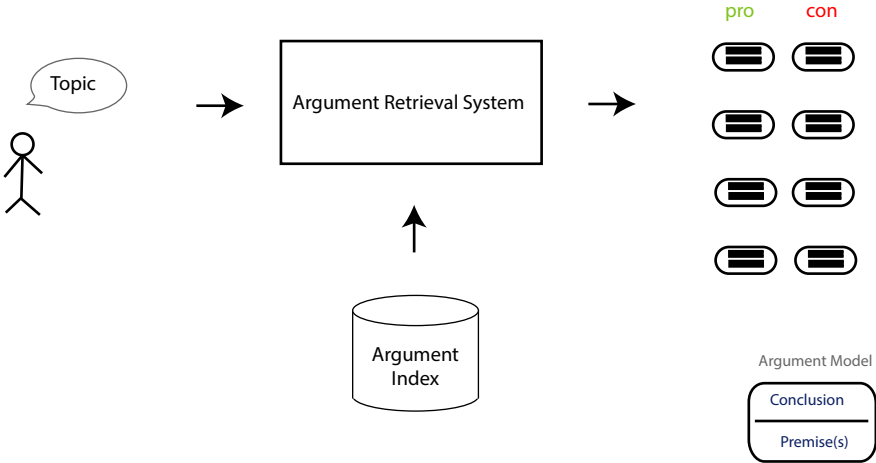
[At 18 we also can make the choice to enroll in the...](#)  
► Show full argument  
At 18 we also can make the choice to enroll in the dangerous life of a soldier. ... Yes, but the state **should** not baby it's citizens through making choices in their life. ...  
<https://www.debate.org/debates/Marijuana-should-remain-illegal/1/> score ▾

[I don't deny this, however what my opponent does not...](#)  
► Show full argument  
I don't deny this, however what my opponent does not address in round 2 is if these bans work or not. ... "This "gift of nature" is a cause of several thousand employment terminations, incarcerations, relationship breakups, car ...  
<https://www.debate.org/debates/Marijuana-should-remain-illegal/1/> score ▾

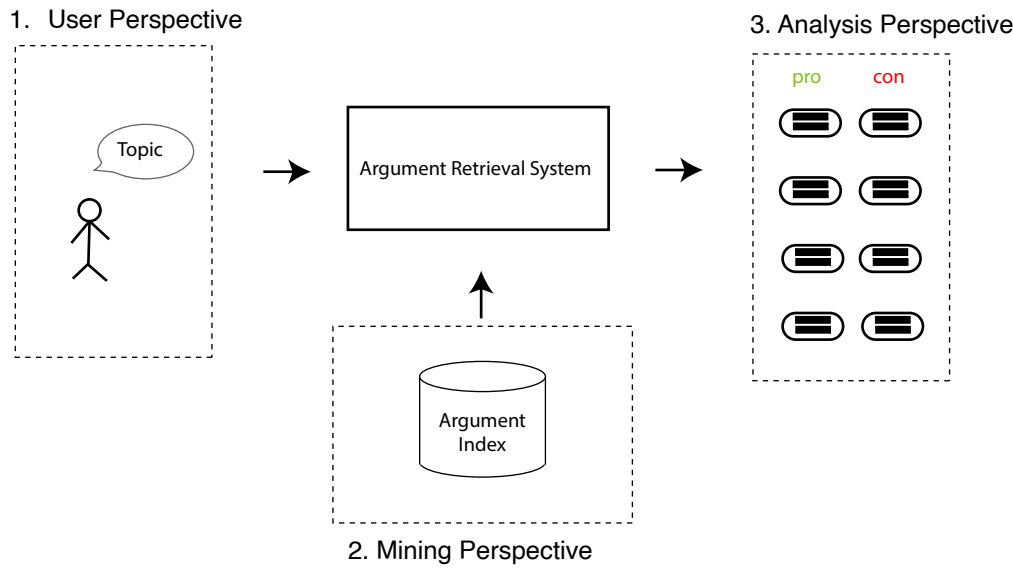
Argument retrieval systems promote:

- ❑ Transparency
- ❑ Explainability

# Contributions



# Contributions

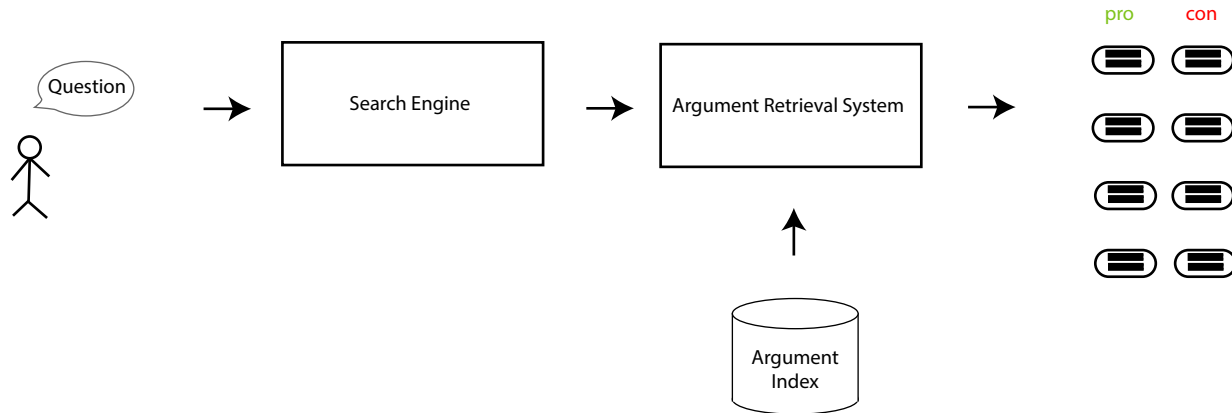


1. Identifying argumentative questions in web search engines logs
2. Assessing topic bias in argument corpora
3. Frame identification of arguments



# Contributions

Goal: Integrating argument retrieval technology in web search engines.



1. Identifying argumentative questions in web search engines logs
2. Assessing topic bias in argument corpora
3. Frame identification of arguments

RQ1. How to identify questions that look for arguments in the query stream of a search engine?

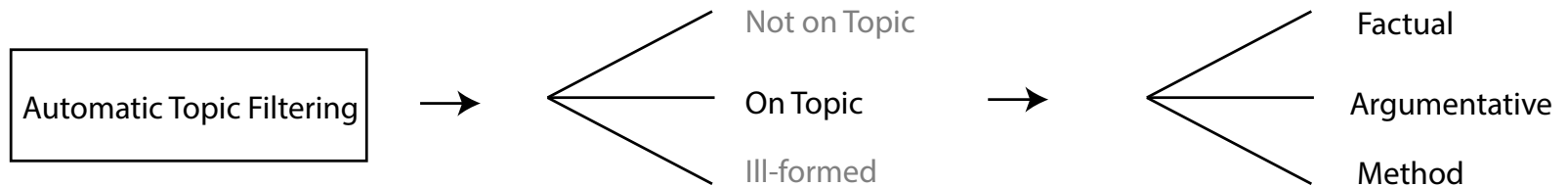
# 1) Identifying Argumentative Questions

## Preparing a Russian Questions Dataset

1. Filter from Yandex logs 4.5 million Russian questions on 19 controversial topic

Example topics: Putin, Navalny, Nord Stream, and marijuana

2. Sample 54,850 questions and annotate them with the annotation scheme:



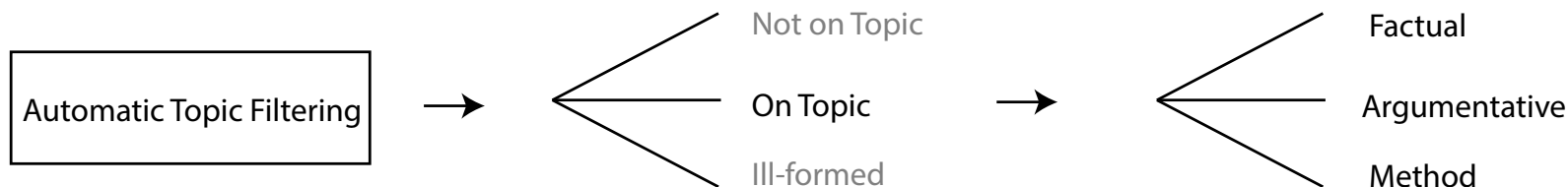
# 1) Identifying Argumentative Questions

## Preparing a Russian Questions Dataset

1. Filter from Yandex logs 4.5 million Russian questions on 19 controversial topic

Example topics: Putin, Navalny, Nord Stream, and marijuana

2. Sample 54,850 questions and annotate them with the annotation scheme:



## Statistics and Examples:

| Question Type | Percentage | Count  | Example                                |
|---------------|------------|--------|----------------------------------------|
| Factual       | 64%        | 25,332 | Is marijuana legalized in Belgium?     |
| Argumentative | 28%        | 10,982 | Will the president legalize marijuana? |
| Method        | 8%         | 3,026  | How to use medical marijuana?          |

# 1) Identifying Argumentative Questions

## Analysis of Questions Characteristics

Comparison of argumentative questions with factual and method questions using lexical and syntactical patterns.<sup>1</sup>

| Question Type | Starts with wh-words (except <i>why</i> ) | Starts with <i>why</i> | Formed as yes/no | Asks for predictions | Asks for comparisons | Subject is personal pronoun | Others |
|---------------|-------------------------------------------|------------------------|------------------|----------------------|----------------------|-----------------------------|--------|
| Factual       | 65.7%                                     | 1.3%                   | 7.2%             | 3.8%                 | 3.2%                 | 0.3%                        | 18.5%  |
| Argumentative | 41.3%                                     | 20.7%                  | 13.8%            | 8.2%                 | 5.7%                 | 3.8%                        | 6.5%   |
| Method        | 93.9%                                     | 0.4%                   | 0.0%             | 0.6%                 | 4.4%                 | 0.4%                        | 0.9%   |

Finding: Argumentative questions look for predictions and explicitly for reasons.

<sup>1</sup>Some question characteristics overlap (e.g., asks for predictions and asks for comparisons.)

# 1) Identifying Argumentative Questions

## Question Type Classification

Developing classifiers to map questions to argumentative, factual or method.

Experimental setting is leave-one-topic-out: test on one topic after training on remaining topics.

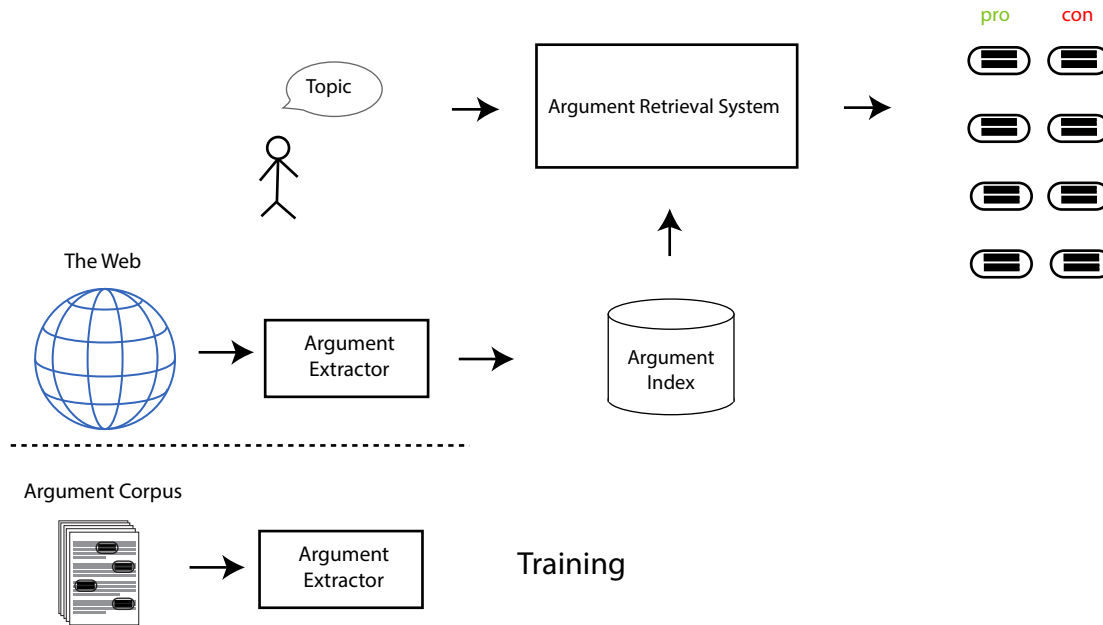
F1-score of the three question types and their macro average.

| <b>Classifier</b>   | <b>Factual</b> | <b>Argumentative</b> | <b>Method</b> | <b>Macro</b> |
|---------------------|----------------|----------------------|---------------|--------------|
| Majority Baseline   | 0.78           | 0.00                 | 0.00          | 0.26         |
| Logistic Regression | 0.80           | 0.61                 | 0.52          | 0.65         |
| RuBERT              | 0.85           | 0.74                 | 0.74          | 0.78         |

Finding: Identifying argumentative questions is feasible, even on unseen topics.

# Contributions

Goal: Fostering the generalizability of argument mining approaches over topic.



1. Identifying argumentative questions in web search engines logs
2. Assessing topic bias in argument corpora
3. Frame identification of arguments

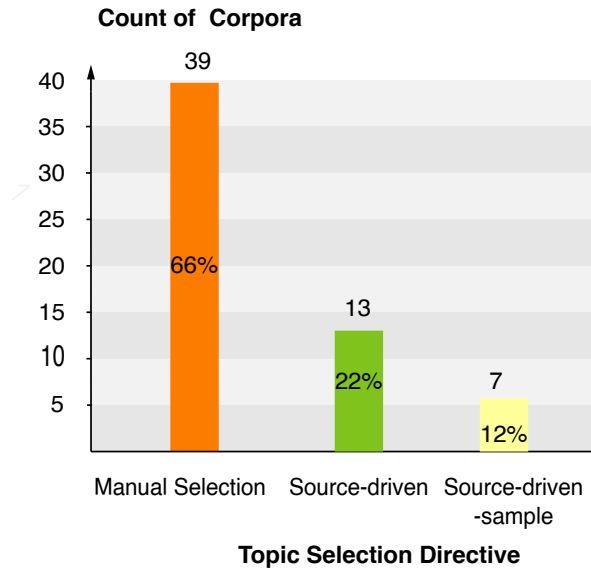
RQ2. How well do argument corpora represent controversial topics?

## 2) Topic Bias in Argument Corpora

### Survey Regarding Topic Selection

A survey of 59 argument corpora shows that researchers take three approaches:

- ❑ Manual selection: choosing a set of topics manually
- ❑ Source-driven-greedy: a whole source is exploited
- ❑ Source-driven-sample: a source is sampled

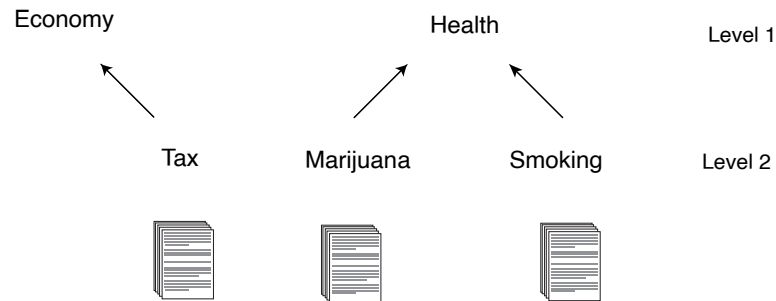


# 2) Topic Bias in Argument Corpora

## Trustworthy Topic Ontologies

Topic ontology: a directed graph where

- Nodes are topics
- Edges indicate is part of relation: topics that are part of other topics are called subtopics.



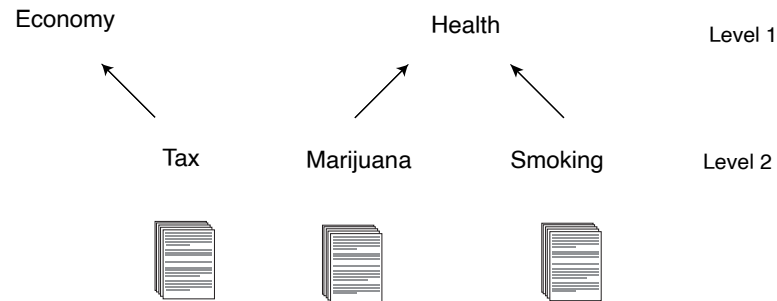


## 2) Topic Bias in Argument Corpora

### Trustworthy Topic Ontologies

Topic ontology: a directed graph where

- Nodes are topics
- Edges indicate is part of relation: topics that are part of other topics are called subtopics.



Three trustworthy topic ontologies with categorized documents

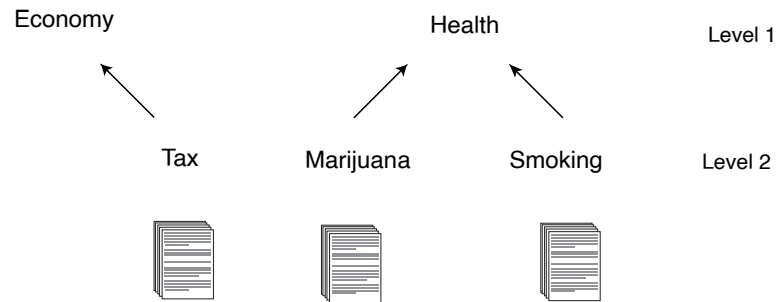
- World Economic Forum (WEF): global issues (mainly economical)
- Debatepedia: biased to western culture
- Wikipedia

# 2) Topic Bias in Argument Corpora

## Trustworthy Topic Ontologies

Topic ontology: a directed graph where

- Nodes are topics
- Edges indicate is part of relation: topics that are part of other topics are called subtopics.



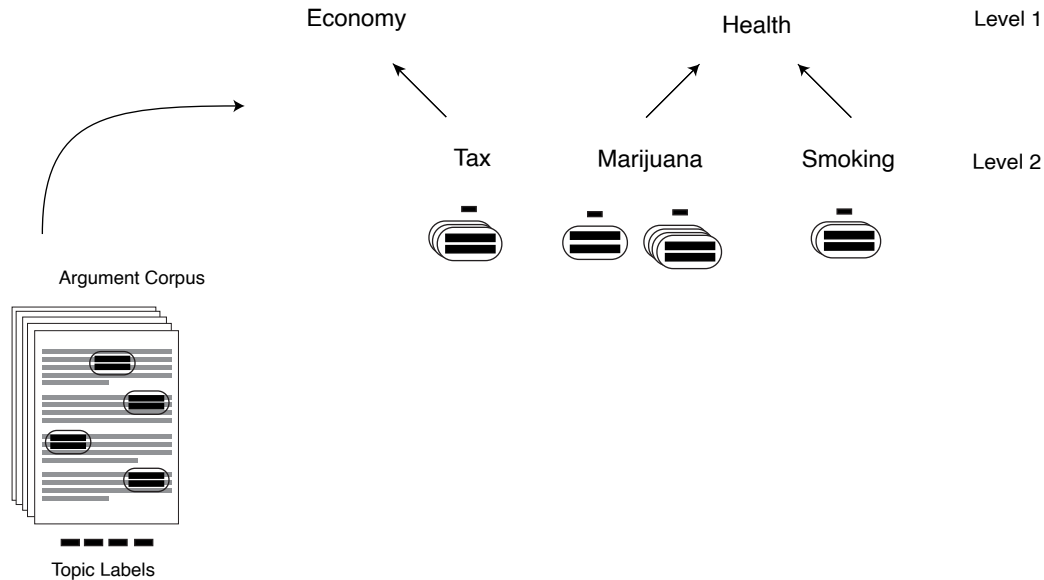
| Ontology                     | Topics | Authors | Docs |
|------------------------------|--------|---------|------|
| World Economic Forum Level-1 | 137    | 334     | 940  |
| World Economic Forum Level-2 | 822    | 217     | 550  |
| Wikipedia Level-1            | 14     | 78,014  | 68   |
| Wikipedia Level-2            | 748    | 1,930   | 1    |
| Debatepedia                  | 89     | 145     | 62   |

# 2) Topic Bias in Argument Corpora

## Units Categorization

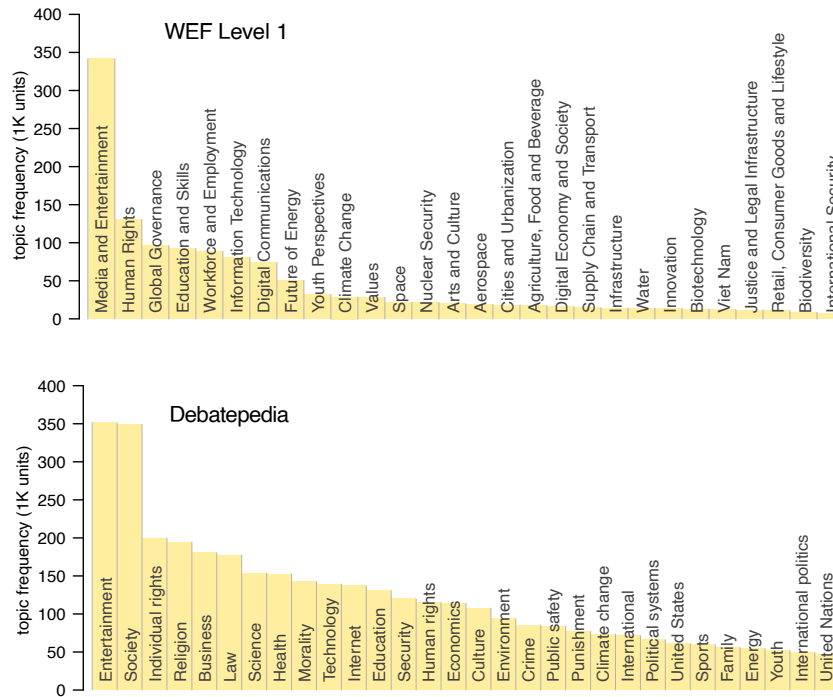
The units of 59 corpora are mapped to the three topic ontologies.

- ❑ Manual:  
mapping the topic labels of a corpus with synonymous or upper topics.
- ❑ Automatic:  
assessing the similarity between a unit and the documents of a topic.



# 2) Topic Bias in Argument Corpora

## Topic Distribution (excerpt)

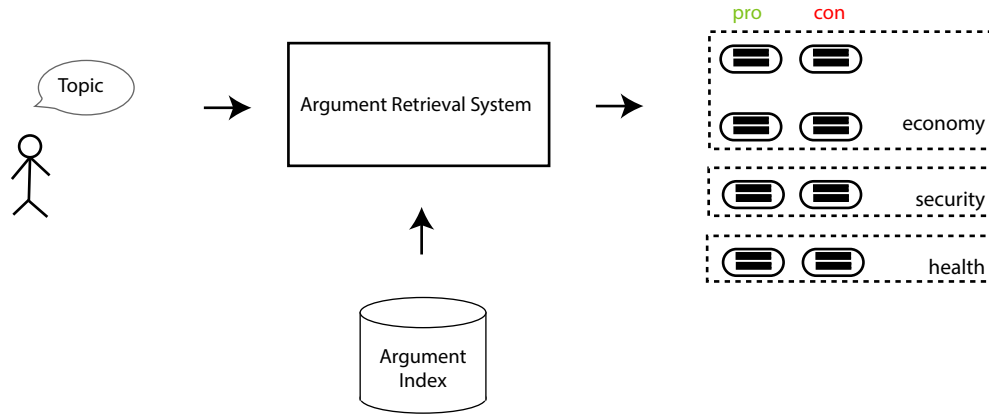


### Findings:

- ❑ The topic distribution of existing argument corpora is skewed and concentrated around a small set of topics.
- ❑ Argument extractors built on these argument corpora might not be generalizable across topics.

# Contributions

Goal: Enable users to select arguments that resonate with their audience.



1. Identifying argumentative questions in web search engines logs
2. Assessing topic bias in argument corpora
3. Frame identification of arguments

RQ3. How to identify the frames of an argument?

# 3) Frame Identification of Arguments

## Introduction

- Framing is to emphasize a specific aspect of a topic while concealing others (Entman et al., 1993).
- A topic like nuclear energy can be framed according to its economical potential or environmental effect among others.



Frame 1: Environment



Frame 2 : Economy

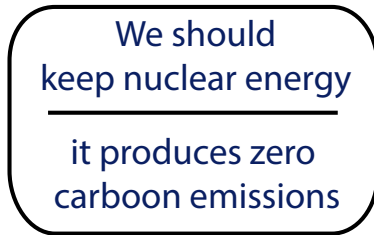
# 3) Frame Identification of Arguments

## Introduction

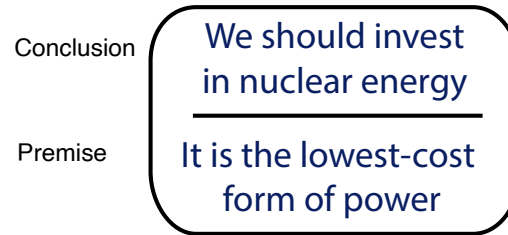
- Framing is to emphasize a specific aspect of a topic while concealing others (Entman et al., 1993).
- A topic like nuclear energy can be framed according to its economical potential or environmental effect among others.

An argument frames a topic by emphasizing an aspect while rejecting others.

Examples:



Frame 1: Environment



Conclusion

Premise

Frame 2: Economy

# 3) Frame Identification of Arguments

## Generic vs Topic-specific Frames

Examples of topic-specific frames:

Bill Clinton  
is a bad president

---

Lewinsky scandal  
lowered his credibility

Frame 1: Lewinsky Scandal

Bill Clinton  
is a good president

---

NAFTA led to  
thousands of jobs

Frame 2: NAFTA

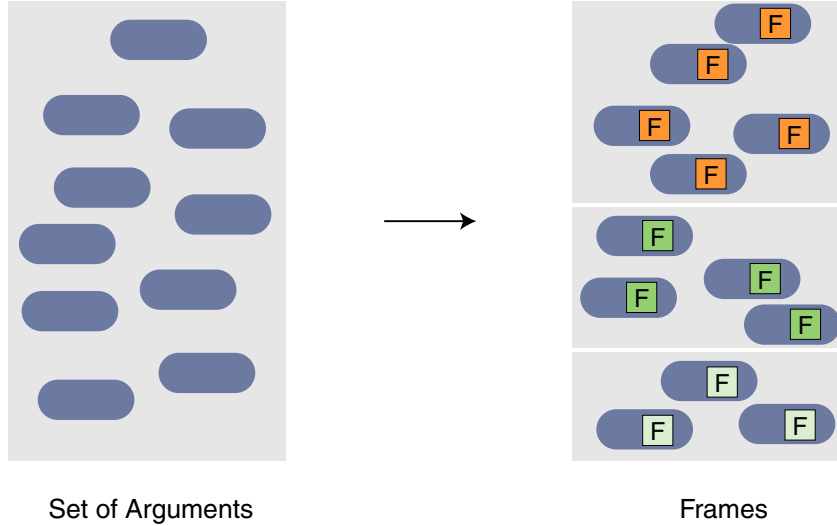
First argument frames dataset covering 467 topics.

| Frame Type     | Count of Frames | Count of Arguments |
|----------------|-----------------|--------------------|
| Generic        | 330             | 7,052              |
| Topic-specific | 1,293           | 5,274              |
| All            | 1,623           | 12,326             |



# 3) Frame Identification of Arguments

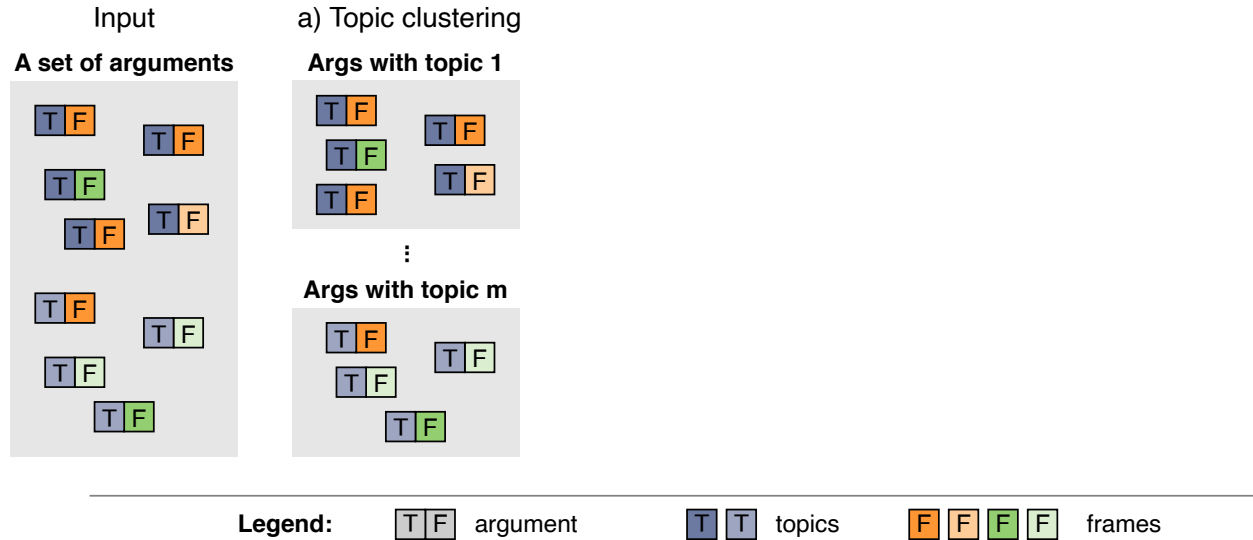
## Approach



- a) Topic Clustering
- b) Topic Removal
- c) Frame Clustering

# 3) Frame Identification of Arguments

Approach: a) Topic Clustering



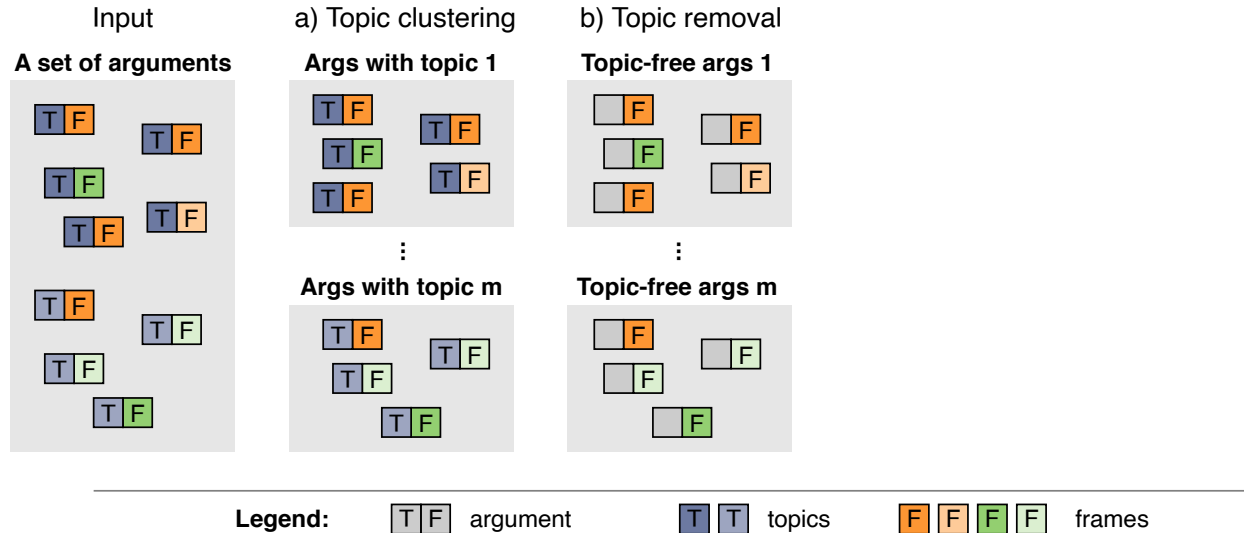
Semantic Spaces:

- TF-IDF
- Latent Semantic Analysis (LSA): a topic model that uses dimension reduction.

Clustering algorithm: K-means with euclidean distance.

# 3) Frame Identification of Arguments

## Approach: b) Topic Removal

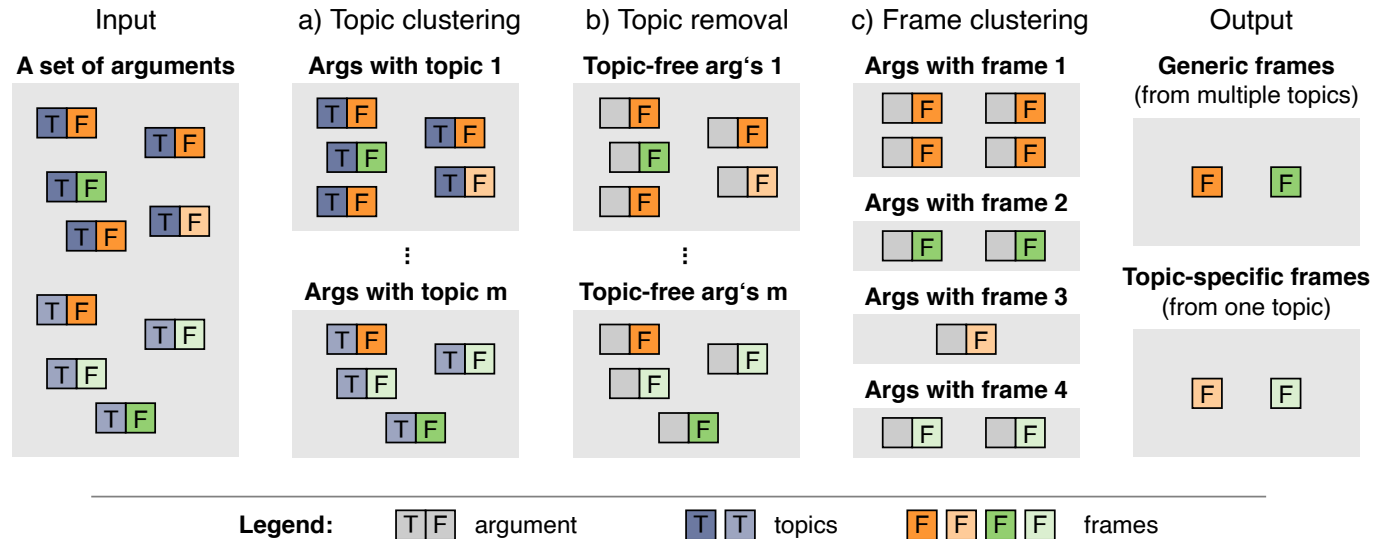


Two models for topic removal:

- Content-based removal:  
Remove tokens with high TF-IDF values in each topic cluster.
- Structure-based removal:  
Remove the conclusion of an argument.

# 3) Frame Identification of Arguments

Approach: c) Frame Clustering



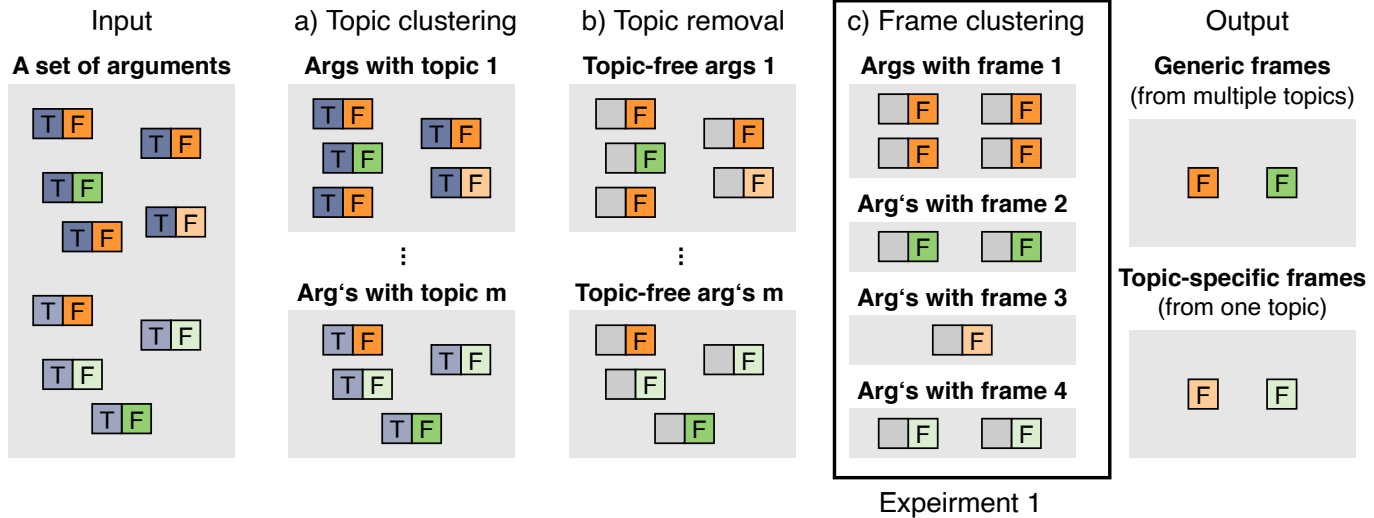
Semantic Spaces:

- TF-IDF
- LSA

Clustering algorithm: K-means with euclidean distance.

# 3) Frame Identification of Arguments

## Experiments



- Topic Clustering
- Frame Clustering

### 3) Frame Identification of Arguments

#### Experiment results: Generic Frame Clustering

Clustering effectiveness in bcubed F1-score.

| Semantic Space | Topic Removal   | Topic Clustering | Frame Clustering |
|----------------|-----------------|------------------|------------------|
| TF-IDF         | No-removal      | 0.45             | 0.19             |
|                | Content-based   | 0.42             | <b>0.28</b>      |
|                | Structure-based | 0.17             | 0.26             |
| LSA            | No-removal      | 0.44             | 0.16             |
|                | Content-based   | 0.40             | 0.21             |
|                | Structure-based | 0.25             | <b>0.20</b>      |

- Removing topic-specific information helps identifying generic frames.
- Structure-based argument removal models is more effective at removing topic-information.

### 3) Frame Identification of Arguments

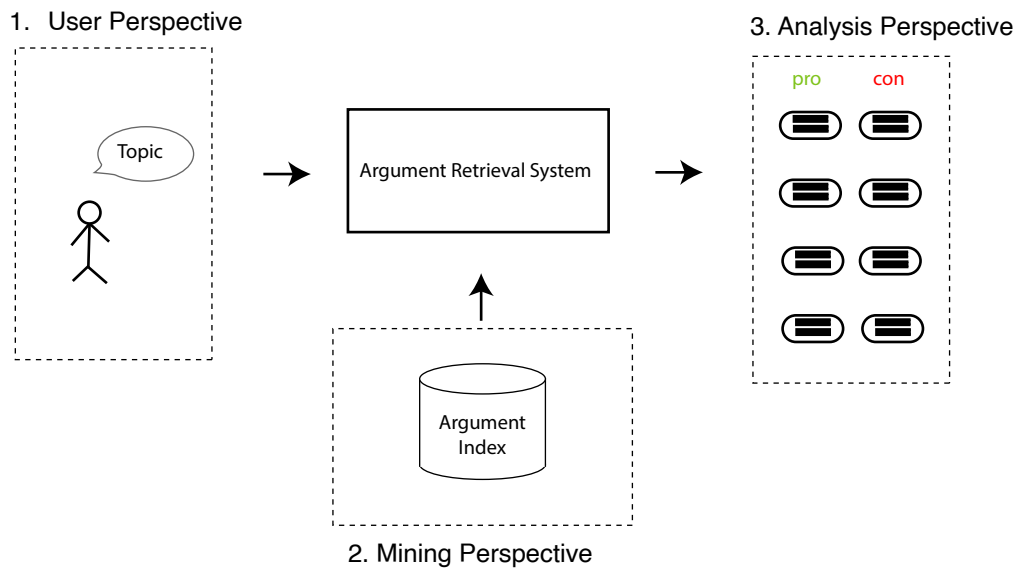
#### Experiment Results: Topic-specific Frame Clustering

Clustering effectiveness in bcubed F1-score.

| Semantic Space | Topic Removal   | Topic Clustering | Frame Clustering |
|----------------|-----------------|------------------|------------------|
| TF-IDF         | No-removal      | 0.45             | <b>0.48</b>      |
|                | Content-based   | 0.42             | 0.45             |
|                | Structure-based | 0.17             | 0.45             |
| LSA            | No-removal      | 0.44             | 0.39             |
|                | Content-based   | 0.40             | <b>0.47</b>      |
|                | Structure-based | 0.25             | 0.46             |

- ❑ Removing topic-specific information helps identifying frames only in LSA space.
- ❑ Using TF-IDF semantic space without topic removal performs the best.

# Conclusion: Research Questions



1. How to identify questions that look for arguments in the query stream of a search engine?
2. How well do argument corpora represent controversial topics?
3. How to identify the frames of an argument?



# Conclusion

## Contributions:

- ❑ Enabling search engines to identify and respond to questions that pertain to controversial topics and those that look for arguments.
- ❑ Method to quantify topic bias in argument corpora and resources to help researchers sample topics in a more representative way.
- ❑ A model and an approach for frames in argumentation.

## Findings:

- ❑ Argumentative questions ask for predictions or reasons.
- ❑ The topic distribution of existing argument corpora is skewed and concentrated around a small set of topics.
- ❑ Identifying the topic of an argument and removing it helps identifying its frames.

# Future Work

- User Perspective
  1. Exploiting session information (i.e., not only one question but a series of questions)
  2. Know more about the user intent (e.g., use case, audience, types of arguments).
  
- Mining Perspective
  1. Developing a unified topic ontology.
  2. Developing topic sampling strategies.
  3. Assessing topic-robustness of argument extractors.
  
- Analysis Perspective
  1. Detecting effective frames sequence from news articles.
  2. Generating frame labels based on argument clusters.

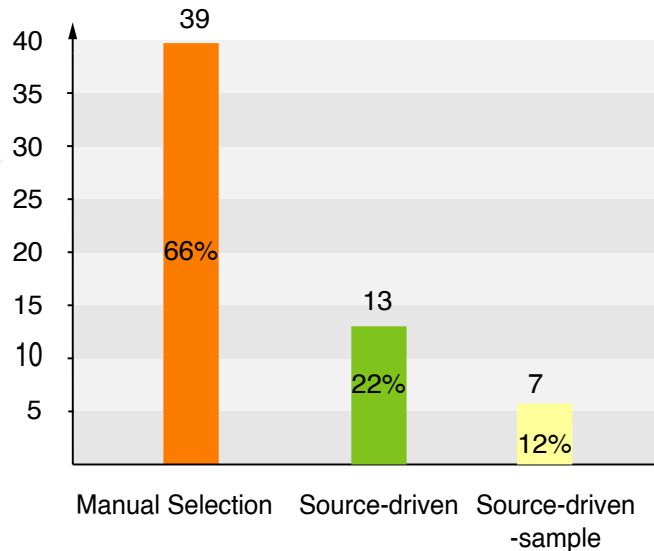
# Topic Bias in Argument Corpora

## Survey Regarding Topic Selection

A survey of 59 argument corpora shows that researchers take three approaches:

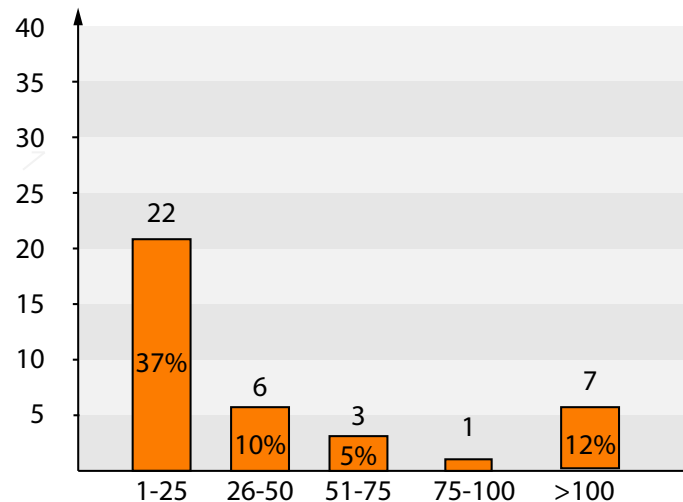
- ❑ Manual selection: choosing a set of topics manually
- ❑ Source-driven-greedy: a whole source is exploited
- ❑ Source-driven-sample: a source is sampled

Count of Corpora



Topic Selection Directive

Count of Corpora



Count of Topic Labels

# Automatic Corpora Unit Categorization

About a third of argument corpora do not provide corpora topic labels and hence is not included in the previous analysis.

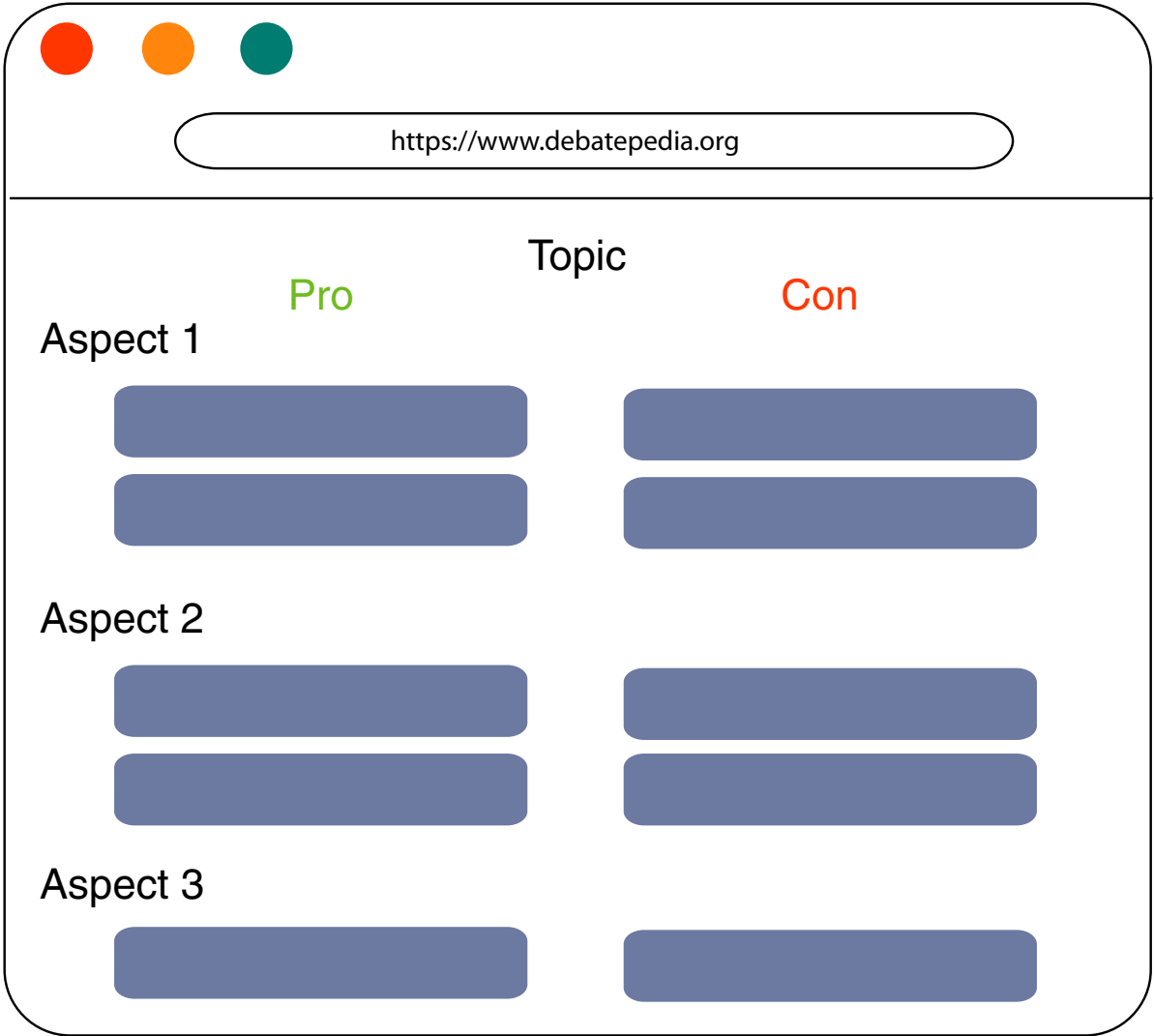
**Approach:** Semantic indexing calculates the cosine similarity between a corpus unit and the documents categorized under an ontology topic.

**Evaluation:** Pooled evaluation for 104 corpora units with a depth of five ontology topics.

F1-score of the approaches

| Approach                    | Wikipedia   |             | WEF         |             |
|-----------------------------|-------------|-------------|-------------|-------------|
|                             | Level-1     | Level-2     | Level-1     | Level-2     |
| Direct match                | 0.06        | 0.40        | 0.29        | 0.19        |
| Semantic Indexing           | 0.43        | <b>0.59</b> | <b>0.34</b> | <b>0.33</b> |
| Text2vec-SI <sub>BERT</sub> | <b>0.47</b> | 0.31        | 0.28        | 0.23        |

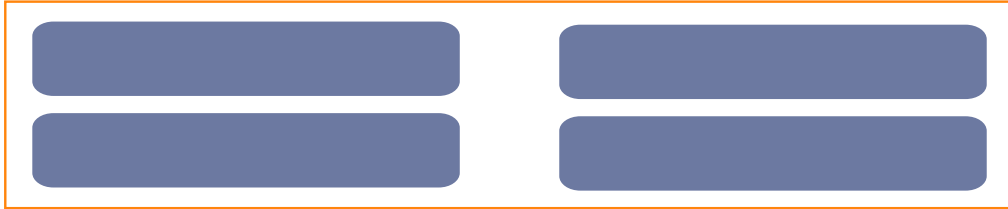
# Dataset Construction from Debatepedia.org



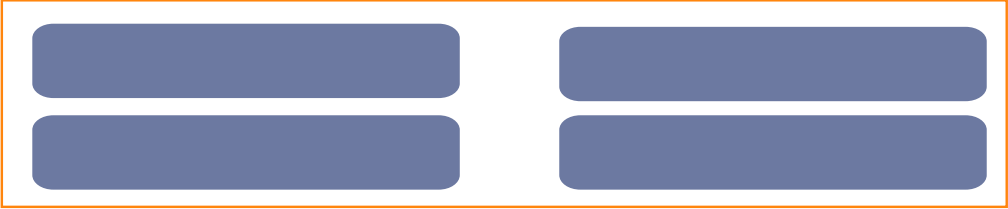
# Dataset Construction from Debatepedia.org

Topic

Frame 1



Frame 2

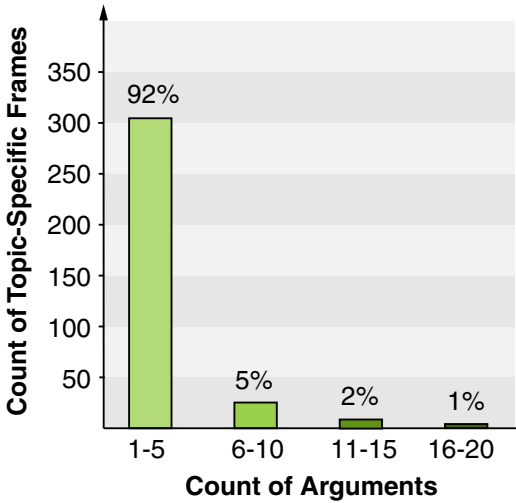
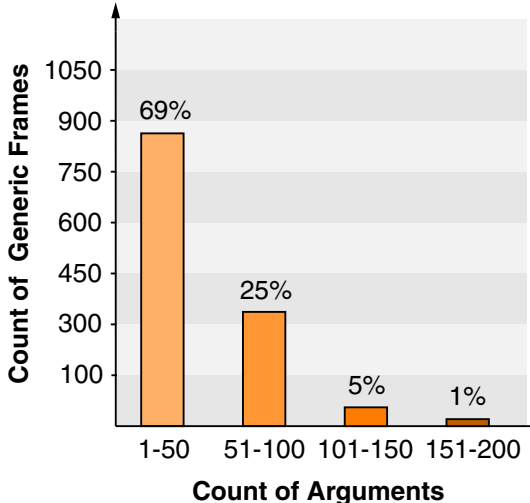
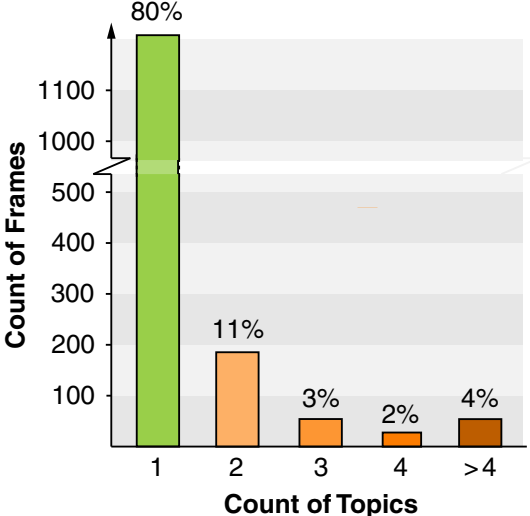


Frame 3



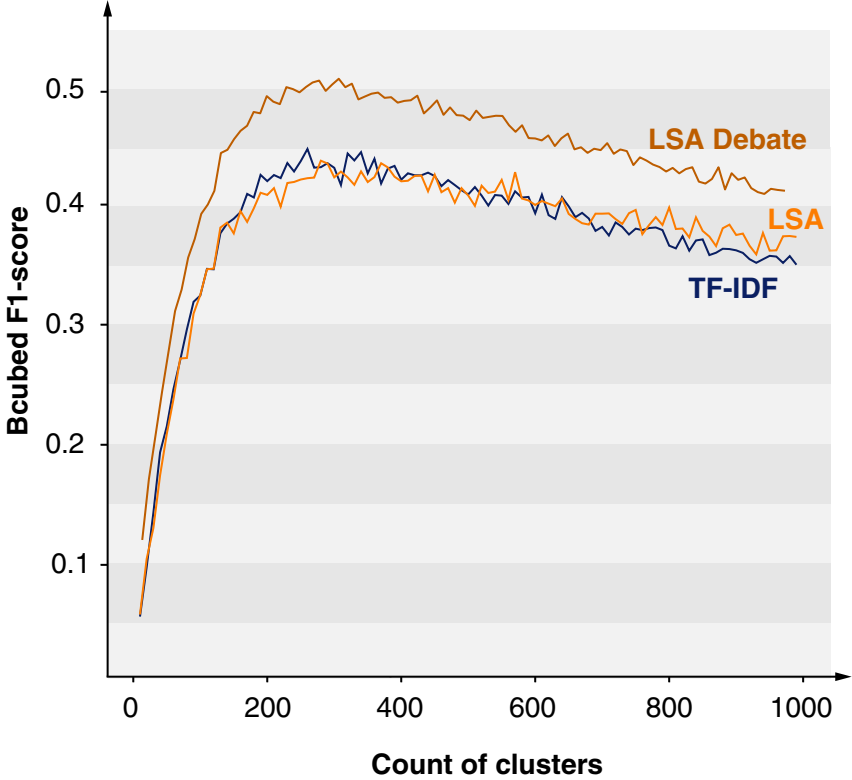
# Dataset Construction from Debatepedia.org

| # Topics | # Frames | # Arguments |
|----------|----------|-------------|
| 467      | 1 623    | 12,326      |



# Experiment Results: Topic Clustering

| Semantic Space    | # Topics   | Bcubed F1   |
|-------------------|------------|-------------|
| <b>LSA Debate</b> | <b>310</b> | <b>0.52</b> |
| TF-IDF            | 260        | 0.45        |
| LSA               | 280        | 0.44        |



LSA Debate is the best semantic space to model the topic of arguments.