Korean Journal of Sport Science 2025, 36(4), 557-567
https://doi.org/10.24985/kjss.2025.36.4.557

Original Article

The Influence of NPMI and TF-IDF-Based Automatic Stopword Generation on Semantic Consistency

Hye-soo Cho[1], Eun-Hyung Cho[2], Hong-suk Kim[3], Soo-Kyung Cho[3], and Ji-Yong Park[3*]

[1]*Department of Sports Science, Hanyang University ERICA*
[2]*Korea Institute of Sports Science*
[3]*Department of Sports Science, Hanyang University*

Article Info
Received 2025. 06. 26.
Revised 2025. 11. 05.
Accepted 2025. 12. 18.

Correspondence*
Ji-Yong Park
ggyy34@hanyang.ac.kr

Key Words
Stopwords,
Topic modeling,
NPMI TF-IDF,
Semantic coherence

This research was supported by the Hanyang University ERICA Industry–University Cooperation Foundation in 2025 (No. 202500000001811).

PURPOSE This study optimized stopword removal to enhance topic modeling performance. We propose an objective method combining normalized pointwise mutual information (NPMI) with median-based term frequency–inverse document frequency (TF–IDF) to automatically generate stopwords. **METHODS** Using text data from 443 research papers on "Taekwondo sparring," we selected stopword candidates based on NPMI and identified 30 words with the lowest TF–IDF scores. We examined the impact of removing 1–30 stopwords on u_mass coherence scores. **RESULTS** The NPMI–TF–IDF method significantly improved coherence (R^2 = .456; p < .001). However, excessive removal led to diminishing returns, with the optimal coherence score (–11.442) achieved at 200 stopwords. In contrast, manually selected stopwords yielded a lower coherence score (–16.001). The findings indicate that integrating TF–IDF with NPMI effectively preserves meaningful words and outperforms PMI^2 and PMI^3 approaches. **CONCLUSIONS** Manual stopword selection can reduce reproducibility. Optimizing stopword removal based on domain-specific characteristics is essential. Future research should validate this method across diverse fields to establish a more generalizable standard.

서 론

내용분석연구는 미디어나 문헌에 내재된 메시지를 체계적으로 분류하여 사회문화적 의미를 탐색하는 방법으로 질적 자료를 객관적이고 체계적으로 양적 자료화할 수 있다는 점에서 학문적으로 높은 활용 가치를 지닌다(Kim, 2005). 그러나 내용분석은 본질적으로 연구자의 해석을 전제로 하기 때문에 동일한 자료로부터 상이한 결론이 도출될 가능성이 존재하며, 이러한 주관성은 연구의 재생가능성과 신뢰도를 저해하는 주요 요인으로 지적되어 왔다.

이러한 한계를 극복하기 위해 최근에는 빅데이터와 머신러닝 기반의 텍스트 마이닝 기법이 도입되었으며, 특히 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA)을 기반으로 한 토픽 모델링이 질적 자료를 통계적으로 구조화할 수 있는 대표적 분석 도구로 자리 잡았다(Blei et al., 2003).

LDA는 문서 집합 내 단어의 공기(共起, co-occurrence) 패턴을 확률적으로 모델링함으로써 잠재 주제를 도출할 수 있다는 점에서 기존의 내용분석보다 높은 객관성을 확보할 수 있다. 이러한 장점으로 인해 체육학 분야에서도 토픽 모델링이 점차 확산되고 있으며, 운동생리학(Lee et al., 2024), 스포츠심리학(Bae et al., 2024), 체육정책(Kim et al., 2024), 체육측정평가(Cho, 2023), 스포츠교육학(Lee & Cho, 2025) 등 다양한 세부 영역에서 적용 사례가 증가하고 있다.

그러나 토픽 모델링 또한 불용어(stopwords) 처리 과정에서 연구자의 판단이 개입되는 구조적 한계를 지니기 때문에 완전한 객관성을 담보하지는 못한다. 동일한 단어라 하더라도 연구 주제나 데이터 특성에 따라 핵심어일 수도 제거 대상일 수도 있기 때문이다. 체육학 연구의 경우 '운동', '선수' 등 특정 단어가 거의 모든 문헌에서 빈번하게 등장하므로 연구자가 임의로 이를 불용어로 지정하는 경우가 많다. 그러나 이러한 단어들은 연구 맥락에 따라 핵심 의미를 포함하기도 하므로, 불용어 처리 기준의 차이가 연구 결과의 일관성을 저하시킬 위험이 존재한다.

이러한 문제는 특히 태권도 연구에서 두드러진다. 태권도는 경기 규칙의 변화, 기술 명칭의 세분화, 채점 방식의 개정 등으로 인해 시

기별·연구자별로 동일한 용어가 상이한 의미로 사용되는 경향이 있다(Jung, 2024; Jang & Bang, 2025). 예를 들어 '공격', '득점', '라운드'와 같은 단어는 경기분석 연구와 기술훈련 연구에서 전혀 다른 맥락으로 활용되며, 이로 인해 단어의 빈도 기반 접근만으로는 주제의 의미를 정확히 파악하기 어렵다. 따라서 태권도 문헌에서 나타나는 특수한 언어적 속성은 불용어 처리의 객관화가 필수적인 과제로 여겨지며, 이는 다른 종목보다도 토픽 모델링의 정제 단계가 연구 결과에 미치는 영향이 더 크게 나타날 수있다.

선행연구에서는 불용어 처리 단계의 객관성을 확보하기 위해 단어 간 점별상호정보량(Pointwise Mutual Information, PMI)을 활용한 자동 불용어 생성 방법을 제안하였다(Church & Hanks, 1990). 그러나 희귀 단어에 높은 PMI 값이 부여될 경우, 주요 핵심어가 불용어로 잘못 분류될 수 있다는 한계가 지적되었다(Croft et al., 2010). 이후 Lee and In(2017)은 PMI 기반 자동 불용어 생성 방식을 적용하여 혼잡도(Perplexity)를 평가한 결과, 기존 표준 불용어 리스트보다 낮은 혼잡도를 보여주었다고 보고하였으나, 기본 PMI가 정규화되지 않아 수치 비교가 어렵다는 제한점을 남겼다. 이에 Role and Nadif(2011)는 단어 간 의미적 연관성을 정교하게 평가할 수 있는 정규화된 NPMI(Normalized PMI)를 제시하였으며, TF-IDF(Term Frequency-Inverse Document Frequency) 역시 단어 중요도를 정량적으로 평가하는 보편적 방식으로 활용되어 왔다(Grün & Hornik, 2011; Manning et al., 2008).

이에 본 연구는 불용어 처리를 연구자의 주관적 선정 방식과 자동화된 산출 방식(NPMI 및 TF-IDF 기반)으로 구분하여, 각 방식이 토픽 모델링의 의미론적 일관성(Coherence score)에 미치는 영향을 규명하고자 한다.

본 연구는 방법론적 측면에서 불용어 자동 생성 기법의 타당성을 검증함과 동시에 태권도 연구 문헌이라는 특수한 도메인에서 해당 기법의 적용 가능성을 실증적으로 탐색한다는 점에서 의의를 가지며 나아가 체육학 분야 텍스트 마이닝 관련 연구에서 연구자의 주관성을 최소화하고 재생 가능한 분석 체계를 확립하기 위한 기초적 근거를 제공하는 데 목적이 있다.

연구방법

연구대상자료

본 연구의 목적을 달성하기 위해 학술연구정보서비스(Research Information Sharing Service; RISS)를 활용하여 '태권도 겨루기'를 키워드로 검색하였으며, 2002년부터 연구 착수 시점인 2024년 12월까지 등재된 443편의 논문을 수집하였다. 분석에는 각 논문의 국문 초록을 대상으로 데이터를 추출하여 활용하였고, 2002년대 초반 초록이 그림파일로 저장된 논문의 경우는 직접 작성하여 데이터 세트를 구축하였다.

또한, 본 연구는 KCI에 공개된 연구논문을 이용하여 데이터를 분석한 것이므로 별도의 기관윤리심의위원회(IRB)의 절차를 진행하지 않았다(American Psychological Association, 2017).

1단계 불용어 후보 선정 예제

〈Eq. 1〉은 NPMI 산출 공식으로, Role and Nadif(2011)가 제안한 기본 PMI를 개선한 방식이며, 단어 간 공기 행렬(co-occurrence matrix)을 보다 신뢰성 있게 분석할 수 있어 특정 단어 쌍이 문서 내에서 함께 등장하는 경향성을 보다 정확히 반영할 수 있다. 특히, 결과에 따르면 NPMI 값이 0에 가까운 단어들은 서로 독립적인 맥락에서 등장할 가능성이 높아 불용어일 가능성이 있다고 보고하였다.

아울러 NPMI는 단어 쌍 간의 상호정보를 정규화하여 단어 간 의미적 관계를 보다 정량적으로 측정하는 방법이기 때문에 단순히 특정 단어의 빈도를 측정하는 TF-IDF와 달리 단어 쌍 간의 관계를 고려하기 때문에 보다 정교한 필터링이 가능하다(Role & Nadif, 2011). 특히, NPMI 값이 -.1 ≤ NPMI ≤ .1인 단어 쌍은 의미적으로 독립적인 단어일 가능성이 높다고 볼 수 있다.

$$NPMI(w1, w2) = \frac{PMI(w1, w2)}{-\log P(w1, w2)} \quad \cdots\cdots \text{Eq. 1.}$$

아래 예제 〈Table 1〉은 특정 도메인을 체육학으로 가정하고 자주 사용되는 7개의 핵심 용어를 기반으로 NPMI 계산을 적용한 행렬 예시를 나타낸 것이다. 결과를 살펴보면 '최대산소섭취량'과 '심폐지구력'은 .323으로 가장 높은 값을 보이며, 이는 두 개념이 밀접한 관계를 가짐을 시사한다. 반면, '최대산소섭취량'과 '근력'은 -.198로 두 단어가 상대적으로 독립적으로 등장할 가능성이 높음을 의미한다. 또한, '근력'과 '무산소운동'은 .224로 두 단어가 일정 수준의 연관성을 가짐을 나타낸다. 이와 달리, '운동부하검사'와 '체성분'은 -.132로 상대적으로 낮은 연관성을 보이며, 이를 통해 〈Table 2〉와 같이 불용어 후보를 선정할 수 있다.

Table 1. Example of NPMI-Based Co-occurrence matrix for key terms in sports science

	x1	x2	x3	x4	x5	x6	x7
x1	0						
x2	.099	0					
x3	.224	-.045	0				
x4	0	.099	-.198	0			
x5	-.154	.099	0	.323	0		
x6	.055	-.078	.224	.154	.154	0	
x7	.224	.198	.055	-.132	0	.099	0

x1: Muscular Strength, x2: Effect, x3: Anaerobic Exercise, x4: Maximal Oxygen Uptake, x5: Cardiorespiratory Endurance, x6: Graded Exercise Test, x7: Body Composition

Table 2. Extraction of stopword candidates based on NPMI

Word Pair	-0.1 ≤ NPMI ≤ 0.1	Stop-word
x4: Maximal Oxygen Uptake x2: Effect	.099	O
x4: Maximal Oxygen Uptake x3: Anaerobic Exercise	-.198	X
x6: Graded Exercise Test x4: Maximal Oxygen Uptake	.154	O
x7: Body Composition x6: Graded Exercise Test	.099	O

Table 3. Final stopword dictionary construction

Included NPMI Stopword Candidates	Reference Word	TF-IDF Value	Threshold Comparision	Stopword Selection
x4: Maximal Oxygen Uptake, x2: Effect	x2 (Effect)	.25	≤ .32	O
x4: Maximal Oxygen Uptake, x3: Anaerobic Exercise	x3 (Anaerobic Exercise)	.85	> .32	X
x4: Maximal Oxygen Uptake, x2: Effect, x6: Graded Exercise Test, x4: Maximal Oxygen Uptake	x4 (Maximal Oxygen Uptake)	.36	> .32	X
x6: Graded Exercise Test, x4: Maximal Oxygen Uptake, x7: Body Composition, x6: Graded Exercise Test	x6 (Graded Exercise Test)	.29	≤ .32	O
x7: Body Composition, x6: Graded Exercise Test	x7 (Body Composition)	.21	≤ .32	O

O = Selected as stopword (TF-IDF ≤ .32); X = Excluded (TF-IDF > .32).

Table 4. Summary of the automatic stopword generation process

Step	Input	Computation	Output
ⓐ Probabilities	Tokenized docs	Windowed co-occurrence, min_count/min_co, Laplace at count level	P(w), P(w1, w2)
ⓑ NPMI	P(w), P(w1, w2)	Normalize PMI into NPMI: PMI / -log(P(w1,w2))	Pair scores
ⓒ Candidates	NPMI scores	Filter [-0.1 ≤ NPMI ≤ 0.1] and collect unique terms	Candidate set
ⓓ Finalization	Raw docs + candidates	TF-IDF(mean) → median cutoff(×multiplier)	Final stopwords

2단계 불용어 선정 예제

기존 PMI 방식이 저빈도 단어에 대해 과도한 값을 부여하는 경향이 있으므로 이를 정규화한 NPMI를 활용하여 보다 균형 잡힌 분석을 수행할 수 있다. 이를 기반으로, 본 연구에서는 NPMI를 활용하여 불용어 후보를 선정하고, 이후 TF-IDF를 종합적으로 고려하여 최종적으로 불용어를 확정하는 2단계 필터링 방식을 적용하였다.

본 연구에서는 NPMI 값이 -.1 ≤ NPMI ≤ .1인 단어 쌍을 의미적으로 독립적인 것으로 간주하고, 해당 단어를 불용어 후보로 선정하였다.

또한, 데이터 내에서 지나치게 낮은 빈도를 가지는 단어들이 노이즈로 작용할 가능성을 줄이기 위해 TF-IDF 중위수 기준을 설정하였다. 이는 단어가 전체 데이터에서 일정 횟수 이상 등장해야 분석에 의미 있는 단어로 간주할 수 있다는 가정에 기반하며, 중위수를 적용하는 TF-IDF 기준과 함께 사용될 때 보다 효과적인 필터링이 가능하다. TF-IDF 중위수 값은 본 연구의 데이터셋에서 .32로 가정하여 이를 기준으로 중위수 이하의 단어를 불용어 후보로 선정한 예시이다(Table 3). x3(무산소운동)와 x4(최대산소섭취량)의 경우 TF-IDF 값이 .85로 중위수를 초과하므로 불용어에서 제외되었으며, x2(미치는 영향)와 x6(운동부하검사)은 TF-IDF 값이 각각 .25와 .29로 기준 이하이므로 불용어로 선정되었다. 또한, x4(최대산소섭취량)의 TF-IDF 값이 .36으로 중위수를 초과하므로 불용어에서 제외되었음이 반영되었다. 이를 통해 NPMI와 TF-IDF 기반의 결합 방식을 활용하여 보다 정교한 불용어 필터링을 수행할 수 있음을 확인할 수 있다.

3단계 자동 불용어 생성 프로세스 예제

〈Table 4〉는 특정 텍스트 데이터셋을 대상으로 불용어를 자동 생성하고 각 단어의 상호정보량(PMI)을 산출·저장하는 과정을 개략적으로 제시한것으로 ⓐ 단어 및 동시출현 확률 계산, ⓑ NPMI 계산, ⓒ 불용어 후보 추출, ⓓ TF-IDF 기반 중위수 필터링을 통한 최종 불용어 확정의 네 단계로 구성된다.

연구 재현성을 위한 구조적 설명을 목적으로한 예시는 부록(Appendix)에 제시되어있다. 이 코드는 연구자의 전체 분석 흐름을 이해하기 위한 설명용 단순화 예시이며 실제 분석에서는 본문에 기술한 바와 같이 NPMI 기반 의미 독립성 필터링과 TF-IDF 중위수 기준 필터링을 순차적으로 결합하여 정교하게 수행되었다. 전체적인 연구의 절차는 〈Table 5〉와 같다.

자료분석

실험의 초기 조건으로 데이터를 분석하기 위한 단위를 단어-문서 관계(Words-Articles Network TF-IDF)로 설정하였고, 하이퍼파라미터인 Alpha와 Beta 값을 각각 최솟값 .01, 최댓값 .01, 간격 .01로 고정하여 각 문서가 소수의 주요 토픽에 집중되면서 각 토픽이 소수의 중요한 단어로 구성되도록 설계하였다.

토픽 모델링에서 추출할 토픽 수를 설정하기 위해 불용어를 삽입하지 않은 상태에서 Coherence score인 u_mass score값을 1~50개까지 증가시키며, 도출된 점수를 확인한 결과 Topic 수 K가 9개일 때 가장 안정적인 성능을 보였다.

이후 최적의 불용어 수를 확인하기 위해 NPMI를 기반으로 불용어 후보를 선정한 후 TF-IDF 중위수를 기준으로 불용어를 생성하여 총 800회의 토픽모델링을 실시한 뒤 u_mass 값을 수집하였다.

이 과정에서 각 실험마다 산출된 9개의 Coherence score는 값의 범위가 크고 일부 실험에서는 극단적으로 높은 혹은 낮은 이상치(outlier)가 발생하는 경향을 보였다. 이에 따라 본 연구에서는

Table 5. Research process

Step	Category	Description
1	Initial Parameter Setting	• Unit of analysis: Words–Articles Network (based on TF-IDF) • Tuning of α and β parameters (min=.01, max=.1, interval=.01)
2	Establishment of Initial u_mass Baseline	• Measurement of u_mass coherence without stopword insertion • Preliminary experiment conducted with topic numbers (K = 1–50) to determine the initial K value
3	Automatic Stopword Candidate Generation	• Application of NPMI + TF-IDF–based automatic stopword generation algorithm • Total of 800 stopword candidates generated
4	Analysis of Automatic Stopword Effect	• Regression analysis of the effect of the number of automatically generated stopwords on u_mass
5	Determination of Optimal Number of Stopwords	• Sequential input of 1–800 automatically generated stopwords and collection of corresponding u_mass values • Identification of the number of stopwords yielding u_mass closest to 0
6	Stopword Comparison	• Topic modeling performed using the optimal number of automatically generated stopwords • Topic modeling performed using manually generated stopwords based on prior research (researcher judgment)
7	Statistical Comparison	• Comparison of u_mass mean values between automatic and manual stopword generation methods • Mann–Whitney U test conducted

Coherence score의 분포의 강건성(robustness)과 안정성을 확보하기 위해 각 실험별 1~9개의 Topic Coherence score의 양 끝단의 최댓값과 최솟값을 제거한 절삭평균(trimmed mean)을 최종 비교 지표로 활용하였다.

자료 처리

본 연구에서 수행된 모든 코딩 작업 및 분석은 Colab 환경에서 수행되었으며, 결과 데이터는 CSV 형식으로 저장되어 Excel에서 후속 처리에 활용되었다.

첫째, NMPI와 TF-IDF 기반으로 자동 생성된 불용어의 수가 의미론적 일관성 지수(u_mass score)에 미치는 영향을 검토하기 위해 OriginPro 2016 소프트웨어를 활용하여 단순 회귀분석(Simple Linear Regression)을 수행하였다.

둘째, 본 연구에서는 NPMI와 TF-IDF 기반으로 자동 생성된 불용어 제거 기법과 연구자의 주관성이 개입된 불용어 제거 기법 간 차이를 비교하기 위하여 Topic K를 7로 고정한 후 각 조건에서 산출된 u_mass score를 종속변수로 활용하였다.

정규성 검정 결과(Shapiro–Wilk: 자동 생성 기법 W=.861, p=.156; 연구자 개입 추출 W=.968, p=.887, Kolmogorov–Smirnov (Lilliefors 보정): 두 조건 모두 p=.200) 두 조건 모두에서 유의수준 .05 기준 정규성 가정은 충족되었다.

그러나 본 자료는 조건별 표본 수가 n=7로 매우 작아 정규성 검정의 검정력이 제한적이며, Coherence score가 상 · 하한을 갖는 bounded data로서 분포 비대칭 및 극단값(outlier)에 민감한 특성을 갖기 때문에 본 연구는 단일 통계 가정에 의존하기보다는 보다 보수적이고 분포 민감하지 않은 비모수 방법인 Mann–Whitney U 검정을 차이 분석 기법으로 적용하였다. 이때 모든 가설 검증을 위한 통계적 유의수준은 α=.05로 설정하였다.

연구결과

불용어의 수가 u_mass score에 미치는 영향

〈Table 6, Fig. 1〉은 NPMI와 TF-IDF를 기준으로 분류된 불용어를 1개부터 30개까지 제거했을 경우 u_mass score에 미치는 영향을 보여주는 회귀분석 결과이다.

회귀모델의 유의성 검정 결과, F=23.49(p<.001)로 통계적으로 유의미하였으며, 결정계수 R^2=.456으로 약 45.6%의 설명력을 보였다. 이는 불용어 수가 토픽 모델의 일관성 지표(u_mass)에 유의한 영향을 미치며, 모델의 품질 변동 중 절반 가까이가 불용어 처리에 의해 설명될 수 있음을 의미한다.

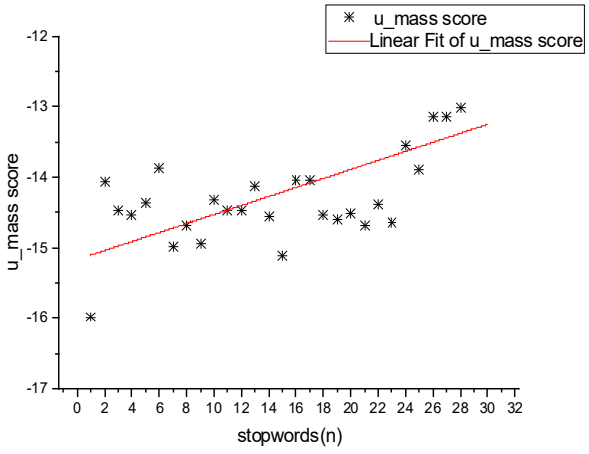

Fig. 1. Coherence score

Table 6. Impact of the number of stopwords on u_mass score

Variable	*B*	SE	*β*	*t*	*F*	R^2
Constant	-15.169	.234		-64.744	23.490***	.456
Stopwords	.064	.013	.675	4.847***		

***p<.001

상수항 -15.169은 불용어가 존재하지 않을 때의 기본 u_mass 수준을 나타내며, 회귀계수 B=.064는 불용어가 1개씩 추가될 때마다 u_mass score가 평균적으로 .064씩 증가함을 보여준다. 이는 불용어가 제거됨에 따라 불필요한 단어의 영향이 줄어들고, 토픽 간 구분도가 점차 개선되어 모델의 의미적 일관성이 높아짐을 나타낸다. 표준 오차는 .013이며, β(표준화된 회귀계수)는 .675로, 불용어 수가 u_mass score에 미치는 효과가 상대적으로 큼을 알 수 있다. 불용어 수의 t-통계량은 4.847로 이 또한 매우 유의한 수준(p<.001)이다.

최적의 불용어 개수 산출 결과

본 연구에서는 토픽 모델링의 의미론적 일관성에 대한 불용어의 영향을 분석하고, 태권도 겨루기와 관련된 토픽모델링 수행을 위한 최적의 불용어 개수를 산출하는 것이 주목적이었다. 이를 위해, 불용어의 수를 1개부터 800개까지 점진적으로 늘려가며 u_mass score의 변화를 관찰하였으며, 결과는 〈Table 7, Fig. 2〉와 같다. u_mass score는 토픽 간의 의미론적 거리를 측정하는 지표로서 0에 가까울수록 더 나은 토픽의 일관성을 의미한다.

분석 결과 불용어 200개를 제거했을 때 u_mass score가 −11.442로 0에 가장 근접한 결과를 나타냈다. 이는 본 연구가 조사한 범위 내에서 최적의 불용어 수로 판단할 수 있다.

Table 7. Changes in u_mass score based on the number of stopwords

Stopwords(n)	u_mass score
30	-15.743
50	-16.762
150	-11.762
200	-11.442
250	-13.844
300	-14.145
350	-15.691
400	-16.283
450	-15.251
500	-15.239
550	-16.578
600	-17.056
650	-18.243
700	-20.685
750	-21.849
800	-22.360

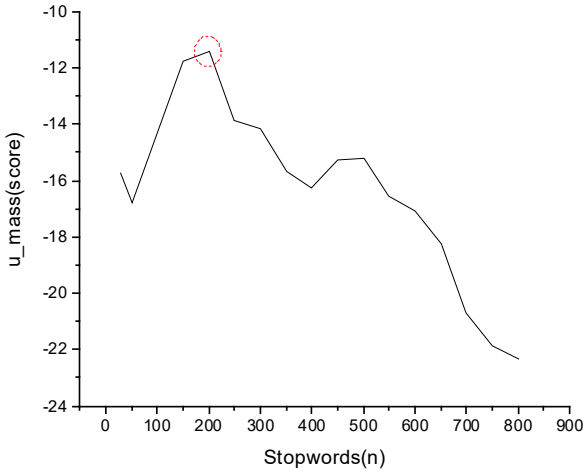

Fig. 2. Changes in u_mass score based on the number of stopwords

불용어 분류 방법 간 평균 u_mass score 차이 분석 결과

불용어 분류 방식에 따른 평균 u_mass score 차이를 비교하기 위해 첫째, NPMI와 TF-IDF를 고려한 자동 생성 불용어의 경우 〈Table 7〉에서 제시된 최적값인 200개를 분석에 활용하였으며, 둘째, 5명의 연구자들 간 합의에 따른 주관적 판단에 의해 선정·제거된 불용어 64개를 분석에 활용하였다. 이때 방법은 독립변수로 방법에 따라 도출된 u_mass score를 종속변수를 통해 Mann-Whitney U-test를 적용한 결과는 아래와 같다.

〈Table 8〉은 분류 방법 간 u_mass score이며, 〈Table 9〉는 두 방법 간 평균과 차이 검정 결과이다. 분석결과 NPMI와 TF-IDF를 고려한 자동 불용어 생성된 u_mass score는 −11.441±1.632이었으며, 연구자들 간 합의에 따라 제거된 u_mass score는 −16.001±.995로 나타나 두 방법 간 유의미한 차이가 있는 것으로 확인되었다(Z=−3.003, p=.003).

방법 간 차이검정 결과를 살펴보면 NPMI와 TF-IDF를 고려한 자

Table 8. Descriptive statistics by classification method

Classification Method	u_mass score
NPMI Classification	-11.23
	-10.16
	-11.46
	-10.01
	-10.17
	-12.57
	-14.5
Subjective Classification	-14.37
	-15.51
	-15.64
	-16.95
	-16.1
	-16.03
	-17.41

Table 9. Comparison of topic coherence (u_mass) between NPMI and subjective stopword classification method

	u_mass score (M±SD)	*Z*	*p*
NPMI Classification (n=7)	-11.442±1.632	-3.003	.003**
Subjective Classification (n=7)	-16.001±0.995		

p<.01**

동 불용어 생성기법에서 최적의 불용어로 선정된 200개를 적용한 u_mass score의 평균이 연구자들이 주관적으로 제거한 불용어에 비해 모든 u_mass score가 0에 가까운 결과를 나타냈다.

논 의

연구는 체육학 문헌에서 자주 등장하는 단어가 맥락에 따라 핵심어 또는 제거 대상이 될 수 있기 때문에 연구자 주관에 의존한 불용어 지정은 결과의 신뢰도와 재생 가능성을 저하시킬 수 있다는 문제와 더불어 태권도 문헌이 시기별 규칙 변화와 용어 사용의 비일관성이라는 고유한 특성을 지니고 있어 일반적인 불용어 처리 방식으로는 토픽 의미가 왜곡될 위험이 높다는 문제 인식에서 출발하였다. 이러한 이유로 본 연구는 태권도 도메인을 통해 검증하지만 향후 체육학 연구에 적용할 수 있도록 재생 가능한 자동 불용어 최적화 기법의 효과를 검증하는데 목적을 두고 있다.

이에 본 연구는 1차적으로 NPMI를 활용하여 불용어 후보 선정 한 뒤 2차로 TF-IDF 중위수 기준으로 불용어를 자동으로 생성한 연구 객관적인 불용어 생성 방법을 제안하였으며, 이를 통해 토픽 모델링을 수행할 때 높은 수준의 의미론적 일관성을 가지는 토픽을 추출할 수 있는 방법을 제시하였다. 이후 NPMI를 활용하여 불용어를 제거한 토픽 모델링과 연구자가 주관적으로 생성한 불용어를 제거한 토픽 모델링 결과를 비교하여 두 방법의 결과에 차이가 있는지를 확인하였다.

먼저, NPMI를 활용하여 불용어를 생성하는 방법이 토픽 모델링 수행에 긍정적인 영향을 미치는지 확인한 결과 이를 위해 '태권도 겨루기'를 키워드로 갖는 443개의 논문에서 추출된 텍스트 데이터에서 TF-IDF 값이 낮은 순서대로 30개의 불용어를 추출하였으며, 불용어 1-30개를 순차적으로 입력하였을 때 u_mass score에 어떤 영향을 미치는지를 검정하였다.

연구결과 R^2=.456(p<.001)로 NPMI와 TF- IDF를 통해 불용어를 제거하는 방법이 의미론적 일관성 점수인 u_mass score의 품질을 향상시키는데 긍정적인 영향을 미치는 것으로 나타났다. 이는 자동 생성 불용어 리스트를 적용하였을 때 불용어가 늘어날수록 토픽 모델 성능 지표인 혼잡도(Perplexity)가 낮아짐을 보고한 선행연구의 결과와도 일치한다(Lee & In, 2017).

이밖에 본 연구의 결과는 Nam and Cheon(2019)이 제안한 PWMH-Gibbs 알고리즘과도 연관이 있다. 해당 연구에서는 LDA 토픽 모델링의 근사추론 과정에서 점별 상호정보량(PMI)을 가중치로 활용하여 불용어를 제거하고 성능을 향상시키는 기법을 제안하였으며, 본 연구에서도 동일한 맥락에서 연구자의 주관성을 배제한 자동 불용어 제거가 토픽 모델의 성능을 개선할 수 있음을 확인하였다.

본 연구의 회귀 분석 결과는 불용어의 개수가 늘어날수록 의미론적 일관성이 높아지는 것처럼 보이나 일정 개수를 초과하면 오히려 효율이 낮아지는 일종의 수확체감의 법칙(diminishing returns)이 나타났다(Schofield et al., 2017). 특히, 불용어 수를 1개부터 30개까지 늘렸을 때 u_mass score가 0에 가까워지는 경향이 나타났다. 이러한 결과로 미루어 볼 때, 불용어가 증가 될수록 u_mass는 0에 가까워져 양호해짐을 알 수 있지만 본 연구에서는 이러한 불용어의 범위를 설정하기 위해 불용어를 800개까지 증가시킨 결과 불용어 200개일 때 u_mass score가 가장 양호한 0에 가깝게 도달했지만 그 이상에 불용어수에서는 오히려 음의 방향으로 이동하며 토픽 품질이 저하되는 역효과가 나타남을 확인하였다.

이는 Schofield et al.(2017)의 연구와 동일하게 과도한 불용어 제거가 오히려 중요 단어까지 제거하는 부작용을 초래하여 모델 성능을 저하시킬 가능성이 있음을 시사한다. 다만, 이러한 결과는 '태권도 겨루기'라는 키워드를 가진 체육학 도메인의 논문을 대상으로 도출된 것이므로 모든 연구에 동일하게 적용할 수 있는 일반적 기준으로 해석하기에는 한계가 있어 타 체육학 연구에서는 불용어 수가 각기 다른 품질을 보일 수 있다. 따라서 특정 도메인에서 이와 같은 연구를 수행할 경우 비정형 데이터의 규모와 특성을 고려하여 최적의 불용어 수를 도출하는 과정이 필요하다고 판단된다.

아울러 본 연구의 핵심인 NPMI를 활용하여 불용어 후보를 선정한 후 TF-IDF 중위수 기준으로 불용어를 자동으로 생성하는 방법과 연구자가 주관적으로 생성한 불용어 리스트를 적용하여 도출된 토픽 모델의 의미론적 일관성 점수가 차이가 있는지를 비교한 결과 자동 생성 불용어를 활용한 경우 토픽 수는 6개로 나타났으며 u_mass score는 -11.442로 기록되었고 의미론적 일관성이 높은 토픽이 도출되었음을 보여주었다.

반면, 주관적으로 생성된 불용어를 활용한 경우, 최적 토픽 수는 5개였으며, u_mass score는 -16.001로 상대적으로 낮았다. 이는 자동 생성 불용어 기법이 연구자가 주관적으로 제거한 방법에 비해 분석과정에 불필요한 단어를 더 효과적으로 제거함으로써 데이터의 노이즈를 줄이고, '태권도 겨루기'라는 도메인에 특화된 키워드를 부각하여 모델 성능을 향상시켰을 것으로 판단된다.

이러한 연구 결과는 저빈도 단어의 점별상호정보량을 다양한 측정 지표와 비교 분석한 Role and Nadif(2011)의 연구결과와 연결지어 논의할 수 있다. 이 연구에서는 PMI 기반의 단어 공기반(co-occurrence) 유사도 측정 방법이 저빈도 단어에 과도한 연관성을 부여하는 문제를 지적하며, 이를 보정하기 위해 정규화된 점별상호정보량(NPMI)을 제안하였다.

다만 NPMI는 0에 가까울수록 의미적으로 독립적인 단어로 간주된다는 장점을 가지지만 저빈도 단어의 경우 정보량 부족으로 인해 불용어 여부 판단이 왜곡될 가능성이 있다. 이를 보완하기 위해 기존 연구에서는 PMI2, PMI3 등 변형 지표의 활용을 제안하였으나, 출현 빈도에 대한 제곱·세제곱 가중 방식으로 인해 오히려 극단값이 과도하게 강조되며 중요한 단어까지 제거될 위험이 존재한다.

이에 본 연구는 단어의 빈도뿐만 아니라 문서 전체에서의 중요도를 반영할 수 있는 TF-IDF를 함께 도입함으로써 의미론적으로 유의미한 단어를 보존하면서 불용어를 보다 정교하게 선별할 수 있도록 설계하였으며, 이를 통해 토픽 모델링의 의미론적 일관성을 향상시키는 실질적 효과를 확인하였다.

이에 따라 본 연구에서는 NPMI에서 불용어 후보로 선정된 단어 쌍

에 Grün and Hornik(2011)가 제안한 TF-IDF 기준 중위수를 적용하였고, 이를 통해 NPMI 값이 낮아 불용어로 분류될 가능성이 높은 단어 중에서도 TF-IDF 값이 높은 단어는 유지할 수 있도록 하여 결과적으로 의미론적 일관성이 향상된 토픽 모델을 구축할 수 있었다.

이러한 결과를 종합했을 때 불용어 제거는 토픽 모델링의 성능을 최적화하는 중요한 과정이며, 무조건적인 불용어 제거가 아닌 최적화된 기준 설정이 필요함을 시사한다. 향후 연구에서는 다양한 도메인에 대한 적용 가능성을 분석하고, 최적의 불용어 개수를 자동으로 결정할 수 있는 기법을 개발하는 방향으로 확장될 필요가 있다. 이를 통해 보다 신뢰성 높은 토픽 모델링을 구현할 수 있을 것이다.

결론 및 제언

본 연구의 결과는 태권도 겨루기 도메인에서 도출된 것이며, 최적 불용어 개수 또한 해당 종목의 언어적 특성과 데이터 구조를 반영한 값이라는 점에서 범용적 기준으로 단정할 수는 없다. 그럼에도 불구하고 본 연구에서 제시한 NPMI–TF-IDF 기반의 불용어 자동 최적화 접근법은 태권도를 포함한 체육학 전반의 텍스트 마이닝 연구에서 토픽 모델링 분석 결과의 품질과 의미론적 일관성을 향상시킬 수 있는 유효한 분석 프레임워크로 확장될 잠재력을 확인했다는 점에 의의가 있다. 이에 향후 연구에서는 태권도 외의 다양한 체육학 도메인에서도 동일한 방식으로 불용어 최적화 실험을 수행하여 본 연구에서 확인된 의미론적 일관성 향상 효과가 일관되게 재현되는지를 검증할 필요가 있다.

나아가 체육학 전반에 적용 가능한 체육학 도메인 기반 불용어 사전을 구축한다면 토픽 모델링 수행 시 연구자의 주관 개입을 최소화하고 분석의 재성가능성을 더욱 강화할 수 있을 것으로 판단된다.

CONFLICT OF INTEREST

논문 작성에 있어서 어떠한 조직으로부터 재정을 포함한 일체의 지원을 받지 않았으며 논문에 영향을 미칠 수 있는 어떠한 관계도 없음을 밝힌다.

AUTHOR CONTRIBUTION

Conceptualization: Hye-soo Cho, Ji-Yong Park, Data curation: Hye-soo Cho, Ji-Yong Park, Formal analysis: Hye-soo Cho, Hong-suk Kim, Methodology: Hye-soo Cho, Soo-Kyung Cho, Project administration: Hong-suk Kim, Ji-Yong Park, Visualization: Soo-Kyung Cho , Eun-Hyung Cho, Writing – original draft: Hong-suk Kim, Ji-Yong Park, Soo-Kyung Cho, Writing – review & editing: Hye-soo Cho, Eun-Hyung Cho

참고문헌

American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct*. https://www.apa.org/ethics/code

Bae, J.-S., Cho, H.-S., Jang, C.-Y., & Song, M.-S. (2024). Analysis of research trends in the Korean Journal of Sport Psychology for the past 20 years (2004 to 2023) using topic modeling. *Korean Society of Sport Psychology, 35*(1), 45-60.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Cho, H.-S. (2023). Natural language processing (NLP) based topic modeling analysis of research trends in the Korean Journal of Measurement and Evaluation in Physical Education and Sport Science. *The Korean Journal of Measurement and Evaluation in Physical Education and Sports Science, 25*(2), 87-102.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, 16*(1), 22-29.

Croft, W. B., Metzler, D., & Strohman, T. (2010). *Search engines: Information retrieval in practice*. Pearson/Addison Wesley.

Grün, B., & Hornik, K. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*(13), 1-30.

Jang, K., & Bang, I. (2025). Exploring the use of Taekwondo technical breaking terms and standardization methods according to competition. *Sport Science, 43*(2), 175-185. 10.46394/ISS.43.2.15

Jung, H. D. (2024). The establishment of correct taekwondo game terminology and educational plans. *The World Society of Taekwondo Culture, 15(*4), 179-190.

Kim, M.-J., Cho, H.-S., & Choi, Y.-S. (2024). Analyzing research trends in the Journal of Korean Society of Sport Policy by applying topic modeling. *Korean Journal of Physical Education, 63*(6), 631-643.

Kim, S. T. (2005). A meta analysis of content analysis research in Korea: Focusing on methodological elements for better content analysis research. *Communication Theories, 1*(2), 39-67.

Lee, H.-K., Lee, D.-W., Cho, H.-S., & Kwak, D. (2024). Analysis of research trends in exercise science using LDA algorithm: Focusing on the period from 2015 to 2024. *Korean Society of Exercise Physiology, 33*(4), 399-408.

Lee, J.-B., & In, H. P. (2017). Automatic generating stopword methods for improving topic model. *In Proceedings of the Annual Conference of Korea Information Processing Society (KIPS)* (pp. 869-872).

Lee, K.-J., & Cho, S.-K. (2025). Analysis of research trends in the Korean Journal of Sport Pedagogy using LDA dynamic topic modeling: Focused on the period from 1994 to 2024. *Korean Journal of Physical Education, 64*(2), 239-257.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Scoring, term weighting and the vector space model. In C. D. Manning, P. Raghavan, & H. Schütze (Eds.), *Introduction to information retrieval* (pp. 100-123). Cambridge University Press.

Nam, S., & Cheon, S. (2019). Inference of latent Dirichlet allocation topic model using PMI. *Journal of The Korean Data Analysis Society, 21*(6), 2789-2800.

Role, F., & Nadif, M. (2011). Handling the impact of low frequency events on co-occurrence based measures of word similarity: A case study of pointwise mutual information. *In Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (KDIR-2011)* (pp. 218-223).

Schofield, A., Magnusson, M., & Mimno, D. (2017). Pulling out the stops: Rethinking stopword removal for topic models. *In Proceedings of the EACL Workshop on Ethics in Natural Language Processing* (April). Valencia, Spain.

NPMI와 TF-IDF를 고려한 자동 불용어 생성 기법이 의미론적 일관성에 미치는 영향: 태권도 연구를 중심으로

조혜수[1], 조은형[2], 김홍석[3], 조수경[3], 박지용[4*]

[1]한양대학교 ERICA 스포츠과학부, 조교수
[2]한국스포츠과학원, 선임연구위원
[3]한양대학교 스포츠과학과, 박사
[4]한양대학교 스포츠과학과, 박사과정

*교신저자: 박지용(ggyy34@hanyang.ac.kr)

[목적] 본 연구는 토픽 모델링의 성능을 향상시키기 위해 불용어 제거 방법을 최적화하는 데 초점을 맞추었다. 이를 위해 정규화된 점별상호정보량(NPMI)과 TF-IDF 기준 중위수를 결합하여 불용어를 자동으로 생성하고 그 효과를 분석하는데 목적이 있다.
[방법] 연구 대상은 '태권도 겨루기'를 키워드로 포함하는 443개의 논문에서 추출한 텍스트 데이터이며 NPMI를 활용하여 불용어 후보를 선정한 후 TF-IDF 값이 낮은 순서대로 30개의 불용어를 추출하였다. 이후 불용어 개수를 1개에서 30개까지 순차적으로 증가시키면서 u_mass score 변화를 측정하였다.
[결과] NPMI와 TF-IDF를 활용한 불용어 자동 생성 방법이 의미론적 일관성 지표(u_mass score)를 향상시키는 데 긍정적인 영향을 미치는 것으로 나타났다(R^2=.456, p<.001). 또한, 불용어 개수가 증가할수록 의미론적 일관성이 향상되는 경향을 보였으나 일정 개수를 초과하면 오히려 성능이 저하되는 수확체감의 법칙(diminishing returns)이 확인되었다. 불용어 200개를 제거했을 때 u_mass score가 가장 높은 값(-11.442)을 기록하였으며 이는 불용어 개수 최적화의 중요성을 시사한다. 반면, 연구자가 주관적으로 선정한 불용어를 적용한 경우 최적 토픽 수는 5개, u_mass score는 -16.001로 상대적으로 낮은 의미론적 일관성을 보였다. 이러한 결과는 NPMI 기반 불용어 제거의 한계를 보완하기 위해 TF-IDF를 함께 고려해야 함을 시사하며 기존 PMI^2, PMI^2 기반 접근법보다 정보량이 높은 단어를 효과적으로 보존할 수 있음을 시사한다.
[결론] 연구자에 의한 주관적 불용어 제거는 연구결과의 재생가능성을 저해하므로 특정 도메인과 데이터 규모에 따라 최적화되어야 할 필요성이 있으며 향후 연구에서는 다양한 도메인에서 본 연구의 방법론을 검증하고 일반화할 필요가 있다.

주요어
불용어, 토픽 모델링, 정규화 점별상호정보량, 의미론적 일관성

※ 이 논문은 2025년 한양대학교 에리카산학협력단의 지원을 받아 수행된 연구임(202500000001811).

[Appendix] Example of an Automatic Stopword Generation Process

ⓐ Term & Co-occurrence Probability Calculation (build_probs)

```
import numpy as np
from collections import Counter, defaultdict

def build_probs(tokenized_docs, window_size=5, min_
count=2, min_co=2):
    term_counts = Counter(t for doc in tokenized_docs
for t in doc)
    term_counts = Counter({t:c for t,c in term_counts.
items() if c >= min_count})
    vocab = set(term_counts)
    total_terms = sum(term_counts.values())
    term_prob = {t: c/total_terms for t,c in term_counts.
items()}

    co_counts = defaultdict(int)
    for doc in tokenized_docs:
        doc = [t for t in doc if t in vocab]
        n = len(doc)
        for i in range(n):
            for j in range(i+1, min(n, i+1+window_size)):
                a,b = sorted((doc[i], doc[j]))
                if a!=b: co_counts[(a,b)] += 1

    co_counts = {k:v for k,v in co_counts.items() if v >=
min_co}
    total_co = sum(co_counts.values())
    co_prob = {k: v/total_co for k,v in co_counts.items()}
if total_co>0 else {}
    return term_prob, co_prob, vocab
```

□ Description:
- tokenized_docs: list of tokenized documents
- window_size: maximum span for co-occurrence window
- min_count: minimum term frequency threshold (terms below this are excluded)
- min_co: minimum co-occurrence frequency threshold (pairs below this are excluded)

□ Operation Flow:
- Compute global term frequency and remove low-frequency terms based on min_count
- Build vocabulary from filtered terms and compute term probability P(w)
- Count co-occurring term pairs within window_size, then exclude pairs below min_co
- Normalize co-occurrence frequencies to compute P(w1, w2)

ⓑ Normalized PMI Calculation (calculate_npmi)

```
def calculate_npmi(w1, w2, term_prob, co_prob,
alpha=0.0, eps=1e-12):
    a,b = sorted((w1,w2))
    pw1 = max(term_prob.get(w1, 0.0), eps)
    pw2 = max(term_prob.get(w2, 0.0), eps)
    p12 = co_prob.get((a,b), 0.0)
    if alpha>0.0:
        p12 = (p12 + alpha) / (1.0 + alpha) # Simple
Laplace Smoothing
    if p12 <= 0.0:
        return 0.0
    pmi = np.log(p12 / (pw1*pw2))
    return pmi / -np.log(p12)
```

□ Description:
- w1, w2: target word pair for NPMI computation
- term_prob, co_prob: dictionaries for term and co-occurrence probabilities
- alpha: Laplace smoothing factor (0.0 = no smoothing)
- eps: small constant to avoid division by zero

□ Operation Flow:
- Retrieve probabilities P(w1), P(w2), P(w1,w2)
- Apply Laplace smoothing to P(w1,w2) if alpha > 0
- If co-occurrence probability is zero → return 0
- Compute PMI(w1,w2) = log(P(w1,w2) / (P(w1)P(w2)))
- Normalize PMI into NPMI: PMI / -log(P(w1,w2))

ⓒ Step 1: Stopword Candidate Extraction (generate_stopwords_candidates)

```
def generate_stopwords_candidates(tokenized_docs,
npmi_low=-0.1, npmi_high=0.1,
    window_size=5, min_count=2, min_co=2, alpha=0.0):
```

```
    term_prob, co_prob, _ = build_probs(tokenized_docs,
window_size, min_count, min_co)
    cand = set()
    for (a,b) in co_prob:
    v = calculate_npmi(a,b,term_prob,co_
prob,alpha=alpha)
        if npmi_low <= v <= npmi_high:
            cand.add(a); cand.add(b)
    return sorted(cand)
```

□ Description:

- npmi_low, npmi_high: lower/upper thresholds for NPMI filtering
- window_size, min_count, min_co, alpha: parameters passed to build_probs and calculate_npmi

□ Operation Flow:

- Call build_probs to obtain term/co-occurrence probabilities
- Compute NPMI for each co-occurring term pair
- Select both terms as stopword candidates if npmi_low ≤ NPMI ≤ npmi_high
- Interpretation: NPMI ≈ 0 implies weak contextual association → low semantic value

ⓓ Step 2: Final Stopword Selection (finalize_stopwords)

```
def finalize_stopwords(raw_docs, npmi_candidates,
median_multiplier=1.0,
                       min_df=2, max_df=0.9, sublinear_
tf=True):
    if not raw_docs or not npmi_candidates: return []
    vec = TfidfVectorizer(lowercase=True, min_df=min_
df, max_df=max_df,
                          sublinear_tf=sublinear_tf)
    X = vec.fit_transform(raw_docs)
    feats = vec.get_feature_names_out()
    mean_scores = np.asarray(X.mean(axis=0)).ravel()
    tfidf = dict(zip(feats, mean_scores))

    cand = [w.lower() for w in npmi_candidates]
    cand_scores = {w: tfidf.get(w, 0.0) for w in cand}
    if not cand_scores: return []

    scores = list(cand_scores.values())
    med = float(np.median(scores))
    thr = med * float(median_multiplier)
    return sorted([w for w,s in cand_scores.items() if s <=
thr])
```

□ Description:

- raw_docs: original (non-tokenized) document list for TF-IDF computation
- npmi_candidates: initial stopword candidates from Step 1
- median_multiplier: scale factor for TF-IDF median threshold
- min_df, max_df, sublinear_tf: TF-IDF vectorizer hyperparameters

□ Operation Flow (NPMI + TF-IDF fusion):

- Compute TF-IDF across entire document set
- Extract average TF-IDF scores for only NPMI-based candidates
- Compute median TF-IDF and apply multiplier to set final cutoff
- Select words whose average TF-IDF ≤ threshold as final stopwords
- Ensures both low contextual value (NPMI) and low discriminative power (TF-IDF) are satisfied