The Impact of Linguistic Features of National Petitions on Policy Agenda-Setting:

A Machine Learning-Based Prediction and Causal Analysis

Seog-Min Lee*

<Abstract>

This study analyzes 427,903 petitions from the Blue House National Petition platform (2017-2019) to examine the impact of linguistic features on policy agenda-setting. Twelve linguistic features were extracted, and machine learning models combined with Propensity Score Matching (PSM) were applied. The Random Forest model achieved an AUC-ROC of 0.91. PSM revealed that text length (ATT=+0.31%p), anger score (ATT=+0.21%p), and specificity score (ATT=+0.13%p) had significant causal effects (all $p<0.001$), while sentiment polarity showed no significant effect. Interaction term logistic regression confirmed significant synergy effects: anger ×specificity (OR=2.126), anger × length (OR=1.605), and specificity × length (OR=1.678), all $p<0.001$. These findings suggest that policy agenda-setting is driven by expression intensity and specificity rather than emotional direction.

Keywords: National Petition, Policy Agenda-Setting, Text Mining, Machine Learning, Propensity Score Matching, Sentiment Analysis

* Professor, Dept. of Public Policy and Big Data Convergence, Hanshin University, Korea

I. 서론

디지털 기술의 발전과 함께 시민과 정부 간 소통 방식이 근본적으로 변화하고 있다. 특히 온라인 청원(e-petition) 제도는 시민들이 직접 정책의제를 제안하고 공론화할 수 있는 새로운 참여 채널로 주목받고 있다. 한국에서는 2017년 8월 청와대(Blue House) 국민청원 제도가 도입되어 약 5년간 운영되었으며, 이 기간 동안 약 60만 건의 청원이 등록되어 시민 참여의 활성화를 보여주었다.

국민청원 제도의 핵심적 특징은 일정 수 이상의 동의를 받은 청원에 대해 정부가 공식 답변을 제공한다는 점이다. 청와대 국민청원의 경우 30일 이내에 20만 명 이상의 동의를 받으면 청와대 수석이나 장관급 인사가 공식 답변을 하도록 설계되어 있다. 이러한 제도적 장치는 시민의 목소리가 정책의제로 전환될 수 있는 공식적 경로를 제공했다는 점에서 의의가 있다.

그러나 모든 청원이 동등한 관심을 받는 것은 아니다. 수십만 건의 청원 중 극히 일부만이 높은 동의 수를 획득하며, 대다수는 소수의 동의에 그친다. 이러한 불균형은 청원의 정책의제화를 결정하는 요인이 무엇인지에 대한 학술적 관심은 부족한 편이다.

기존 연구들은 국민청원 데이터를 활용하여 청원의 주제, 시기별 추이, 동의 분포 등을 분석해왔다(Song & Park, 2019; Eom, 2019). 그러나 이러한 연구들은 대부분 기술적 분석(descriptive analysis)에 그치거나, 청원의 성공 요인을 상관관계 수준에서만 분석하는 한계를 보였다.

특히 다음과 같은 연구의 한계가 존재한다. 첫째, 청원 텍스트의 언어적 특성이 성공에 미치는 영향에 대한 체계적 분석이 부족하다. 텍스트 마이닝 연구가 수행되었으나, 대부분 전통적인 TF-IDF나 키워드 기반 방법에 의존하였다. 둘째, 기존 연구들은 상관관계만을 보고하여, 언어적 특성과 정책 의제화 간의 인과관계를 규명하지 못하였다. 셋째, 언어적 특성 간 상호작용 효과에 대한 분석이 부재하다.

본 연구는 청원 텍스트의 언어적 특성이 정책 의제화에 미치는 영향을 실증적으로 분석하는 것을 목적으로 한다. 구체적으로, 첫째 청원 텍스트의 언어적 특성(감정 강도, 구체성, 길이)이

정책 의제화를 예측할 수 있는지 검토하고, 둘째 특정 언어적 특성이 정책 의제화에 인과적 영향을 미치는지 분석하며, 셋째 언어적 특성들 간 상호작용 효과가 존재하는지 탐색하고자 한다.

선행연구 검토를 바탕으로 다음과 같은 연구 가설을 설정하였다. 가설 1은 분노 표현이 강한 청원이 정책 의제화 확률이 높을 것이라는 것이다. 가설 2는 구체적 정보를 포함한 청원이 정책 의제화 확률이 높을 것이라는 것이며, 가설 3은 텍스트 길이가 긴 청원이 정책 의제화 확률이 높을 것이라는 것이다. 가설 4는 분노, 구체성, 길이가 동시에 높을 때 시너지 효과가 발생할 것이라는 것이다.

본 연구는 방법론적으로는 대규모 한국어 청원 텍스트(약 43만 건)에 머신러닝 기반 텍스트 분석과 성향점수매칭(Propensity Score Matching; PSM)을 결합하여 예측과 인과추론을 동시에 수행하는 분석 프레임워크를 제시하며, 다양한 클래스 불균형 처리 기법(언더샘플링, SMOTE)을 비교하여 심한 불균형 상황에서의 최적 전략을 제시할 것이다. 이론적으로는 정책의제 설정 이론을 디지털 플랫폼 맥락으로 확장하고, 언어적 특성 간 상호작용 효과를 규명할 것이다. 실무적으로는 시민 청원 작성 전략 및 정부의 청원 관리에 대한 정책적 시사점을 도출할 것이다.

II. 이론적 배경 및 선행연구

1. 정책의제 설정 이론

정책의제 설정(policy agenda setting)은 사회 문제가 정부의 공식적인 논의 대상으로 채택되는 과정을 의미한다. Cobb & Elder(1972)는 의제 설정 과정을 공공의제(public agenda)에서 정부의제(governmental agenda)로의 전환으로 개념화하였다. 이 과정에서 문제의 프레이밍, 관심 동원, 정치적 기회가 핵심적 역할을 수행한다.

Kingdon(1984)의 다중흐름모형(Multiple Streams Framework)은 의제 설정을 문제, 정치, 정책의 세 흐름이 만나는 정책의 창(policy window)으로 설명하였다. 이 관점에서

국민청원은 시민이 직접 문제를 정의하고 관심을 동원할 수 있는 제도적 창구로 기능한다.

디지털 환경에서 의제 설정 과정은 전통적 이론과 차별화되는 특성을 보인다. 온라인 플랫폼은 진입 장벽을 낮추고 확산 속도를 높임으로써, 시민 개인이 의제 설정에 직접 참여할 수 있는 기회를 확대하였다(Wright, 2016). 그러나 동시에 관심 경쟁(attention competition)이 심화되어, 성공적인 의제화를 위해서는 효과적인 메시지 구성이 중요해졌다.

2. 전자청원과 시민참여 연구

전자청원(e-petition)에 관한 국제 연구들은 청원의 성공 요인을 다양한 관점에서 분석해왔다. Hagen et al.(2016)는 영국 의회 청원 데이터를 분석하여 언어적, 의미적 특성이 청원 인기도(서명 수)에 영향을 미침을 발견하였다. 특히 감정적 요소, 특히 분노(anger)가 정치 참여를 촉진하는 것으로 나타났다. Chen et al.(2023)은 중국의 온라인 시민 청원을 분석하여 청원의 표현 방식이 정부 응답에 영향을 미침을 보여주었다. 이들은 텍스트 분석을 통해 정부 응답 여부를 예측할 수 있음을 실증하였다.

국내 연구에서는 청와대 국민청원에 대한 분석이 있다. Song and Park(2019)은 자연어처리 기법을 활용하여 청원 분석을 수행한 결과, 다수의 글이 반드시 다수의 동의로 이어지지 않음을 발견하였다. 또한 슬픔보다 분노가 동의 획득에 더 효과적이며, 분노는 심각한 갈등적 의제보다 일상적 의제와 결합할 때 더 큰 호응을 얻는다고 보고하였다.

Eom(2019)은 빅데이터 분석을 통해 동의 양상이 전형적인 멱함수 분포(power-law distribution)를 따르며, 극소수 청원만이 폭발적 동의수 증가를 보임을 발견하였다. 또한 소셜미디어와 언론보도가 확산에 핵심적 역할을 함을 강조하였다.

3. 텍스트 분석과 감성분석

텍스트의 감성(sentiment)이 온라인 콘텐츠의 확산에 미치는 영향은 다양한 분야에서 연구되어 왔다. 일반적으로 감정적 강도가 높은 콘텐츠가 더 널리 공유되며, 특히 분노나 놀람 같은 고각성(high-arousal) 감정이 확산에 효과적인 것으로 알려져 있다(Berger & Milkman, 2012).

한국어 텍스트 분석에서는 사전훈련 언어모델의 적용이 증가하고 있는 추세이다. BERT(Devlin et al., 2019)를 기반으로 한 KoBERT는 SKTBrain(2019)이 개발한 한국어 특화 모델로, 한국어 위키피디아와 뉴스 데이터로 사전학습되어 다양한 하위 작업에 높은 성능을 보인다. 하지만 정책 분야, 특히 국민청원 분석에 KoBERT를 본격적으로 적용한 연구는 아직 부족한 실정이다.

기존 연구들의 검토를 통해 첫째, 언어적 특성과 성공 간의 인과관계를 엄밀히 검증한 연구가 부재함을 알 수가 있다. 둘째, 복수의 언어적 특성이 결합될 때의 시너지 효과를 분석한 연구가 드물다는 것이다. 셋째, 청원의 정책의제화를 예측하는 모델 개발 연구는 없는 실정이다. 넷째, 전체 청원 데이터를 활용한 체계적 분석이 부족하다. 그래서 본 연구는 약 43만 건의 청원 전수 데이터를 활용하고, 머신러닝 예측과 PSM 인과분석을 결합하며, 상호작용 효과를 분석함으로써 이러한 연구의 공백을 해소하고자 한다.

III. 연구 방법

1. 데이터

연구를 위해 GitHub에 공개된 청와대 국민청원 아카이브(Lovit, 2019)를 활용하였다. 이 데이터셋은 2017년 8월부터 2019년 8월까지 등록된 청원을 포함하며, 원본 데이터는 436,660건이다. 데이터 전처리 과정에서 다음의 청원은 제외하였다: (1) 텍스트가 비어있는 청원, (2) 제목과 본문을 합쳐 50자 미만인 청원, (3) 중복 청원(동일 제목 및 본문). 또한 텍스트 정규화 과정에서 특수문자 제거, URL 제거, 연속 공백 정리를 수행하였다. 최종 분석 데이터는 427,903건이다. 분석의 타당성 확보를 위해 데이터를 학습(train), 검증(validation), 테스트(test) 세트로 분할하였고, 종속변수의 클래스 비율을 유지하기 위해 층화 샘플링(stratified sampling)을 적용하였다.

Table 1. Data Split Summary

Split	N	Proportion	Positive Rate
Train	299,532	70%	0.21%
Validation	64,185	15%	0.21%
Test	64,186	15%	0.21%
Total	427,903	100%	0.21%

2. 변수 측정

1) 종속변수: 정책 의제화

본 연구는 청원의 정책의제화를 10,000명 이상의 동의를 받은 경우로 조작적 정의하였다. 청와대 공식 답변 기준(200,000명)을 적용할 경우 양성 사례가 65건에 불과하여 통계적 분석이 어렵기 때문에, 선행연구(Song & Park, 2019)를 참고하여 10,000명 기준을 채택하였다.

$$\text{policy\_agenda} = 1 \quad \text{if} \quad \text{votes} \geq 10{,}000$$
$$= 0 \quad \text{otherwise}$$

전체 데이터에서 양성 비율은 0.21%(약 900건)로 극심한 클래스 불균형이 존재한다. 10,000명 기준은 청와대 공식 기준(200,000명)과 차이가 있으나, 다음과 같은 근거에서 타당하다고 판단하였다. 첫째, 전체 청원의 상위 0.21%에 해당하여 실질적인 대중적 관심을 획득한 청원으로 볼 수 있다. 둘째, 10,000명 이상 동의를 받은 청원은 언론 보도 및 정치적 논의 대상이 되는 경우가 많아 의제화의 실질적 의미를 갖는다. 셋째, 민감도 분석 결과 5,000명, 20,000명 임계값에서 PSM 분석을 통해 주요 변수의 효과 방향과 통계적 유의성이 일관됨을 확인하였으며, 50,000명 이상 임계값에서는 양성 표본 수 제약으로 PSM 대신 t-검정을 적용하여 동일한 방향성을 확인하였다(Table 11 참조).

2) 독립변수: 언어적 특성

8개의 언어적 특성 변수(핵심 독립변수)와 4개의 통제변수를 추출하였다.

Table 2. Variable Definitions

Panel A: Key Independent Variables (Linguistic Features)

Variable	Description	Measurement
text_length	Text length (characters)	len(text)
word_count	Word count	Space-delimited tokens
sentence_count	Sentence count	Period/question/exclamation marks
avg_sentence_length	Average sentence length	text_length / sentence_count
avg_word_length	Average word length	Mean characters per word
sentiment_score	Sentiment score (-1 to +1)	Dictionary-based
anger_score	Anger score (0 to 1)	Anger keyword frequency
specificity_score	Specificity score (0 to 1)	Ratio of numbers/dates/institutions

Panel B: Control Variables

Variable	Description	Measurement	Role in Analysis
category_encoded	Category code	17-category encoding	PSM covariate
topic_id	Topic ID	Category-based mapping	PSM covariate
year	Year	Petition registration year	PSM covariate
month	Month	Petition registration month	PSM covariate

통제변수 4개(category_encoded, topic_id, year, month)는 청원의 주제, 시기적 특성이 정책의제화에 미치는 혼란효과를 통제하기 위해 투입하였다. 특정 카테고리(예: 인권/성평등, 정치개혁)가 시기별로 상이한 사회적 관심을 받을 수 있으며, 연도와 월은 계절적 효과

및 정치적 시의성을 반영한다. 이 변수들은 머신러닝 예측 모델에는 포함하지 않았으며, PSM 분석에서 성향점수 추정 시 공변량으로만 활용하였다. 이는 본 연구의 핵심 관심사가 언어적 특성의 예측력이므로, 맥락 변수를 예측 모델에 포함할 경우 언어적 특성의 독립적 기여도를 과소추정할 수 있기 때문이다. 한편, PSM에서는 처치–통제 간 교란 요인을 최대한 통제해야 하므로 통제변수를 공변량에 포함하였다.

3) 핵심 변수 산출 공식

감성 점수(sentiment score)는 긍정 키워드 15개(좋, 행복, 감사, 희망 등)와 부정 키워드 23개(문제, 피해, 불법, 범죄 등)로 구성된 감성 사전을 활용하여 다음과 같이 산출하였다. 여기서 N_positive는 긍정 키워드 출현 횟수, N_negative는 부정 키워드 출현 횟수이다. 분모에 1을 더하여 키워드가 없는 경우 0을 반환하도록 처리하였다.

$$\text{sentiment\_score} = (\text{N\_positive} - \text{N\_negative}) / (\text{N\_positive} + \text{N\_negative} + 1)$$

아래의 분노 점수(anger score)는 아래의 식과 같이 분노 관련 키워드 12개(분노, 화나, 처벌, 엄벌, 구속, 징역 등)의 출현 빈도를 정규화하여 측정하였다. 여기서 N_anger는 분노 키워드 출현 횟수이다. 분노 키워드는 전체 단어 수 대비 희소하게 출현하므로, 계수 10은 점수의 변별력을 확보하기 위한 보정 계수이다.

$$\text{anger\_score} = (\text{N\_anger} / \text{word\_count}) \times 10$$

아래의 구체성 점수(specificity score)는 구체적 정보 요소의 비율로 산출하였다. 여기서 N_numeric은 숫자를 포함한 토큰 수, N_date는 날짜 패턴 수, N_organization은 기관명 패턴 수이다

$$\text{specificity\_score} = (\text{N\_numeric} + \text{N\_date} + \text{N\_organization}) / \text{word\_count}$$

4) 사전 기반 감성분석의 타당성

사전 기반(dictionary-based) 방법으로 감성과 분노를 측정하였다. 이 방법은 KoBERT 등 사전훈련 언어모델 대비 맥락 이해에 한계가 있으나, 다음과 같은 근거에서 본 연구 맥락에 타당하다고 판단하였다. 청원 텍스트의 특성상 감정 표현은 직접적이고 명시적이다. 청원자는 문제의 심각성을 호소하기 위해 분노, 처벌, 엄벌 등의 명시적 표현을 사용하는 경향이 있어, 키워드 기반 방법의 효용성이 높다. 또한 선행연구에서 유사한 방법이 유효함을 입증하였다. Hagen et. al.(2016)는 영국 청원 분석에서 키워드 기반 감성 측정이 서명 수를 유의하게 예측함을 보였으며, Song and Park(2019)도 동일한 접근으로 유의미한 결과를 도출하였다.

연구의 목적은 감성의 정밀한 측정보다 상대적 강도 비교이다. 처치집단(고분노)과 통제집단(저분노)의 구분에는 정밀도보다 일관성이 중요하며, 사전 기반 방법은 이를 충족한다. 다만 이 방법은 부정문, 아이러니, 맥락적 의미를 충분히 반영하지 못하는 한계가 있다. 향후 연구에서 KoBERT 등 사전훈련 언어모델을 적용한 비교 검증이 필요하다.

3. 분석 방법

1) 머신러닝 분류 모델

정책의제화 예측을 위해 세 가지 머신러닝 모델을 학습하였다. 로지스틱 회귀(Logistic Regression)는 해석가능성이 높은 선형 분류기로, 클래스 가중치(class_weight='balanced')를 적용하였으며 정규화 파라미터 C=1.0, 최대 반복수 1000회로 설정하였다. 정규화 파라미터 C는 scikit-learn의 기본값(1.0)을 채택하였으며, 이는 L2 정규화의 강도를 결정하는 역정규화 계수로서, 기본값은 정규화와 적합도 간 균형을 제공한다. 검증 셋에서 C={0.01, 0.1, 1.0, 10.0}을 비교한 결과 AUC-ROC의 차이가 0.01 이내로 미미하여 기본값을 유지하였다.

랜덤 포레스트(Random Forest)는 100개 의사결정나무의 앙상블 모델로, 클래스 가중치를 적용하였으며 최대 깊이 제한 없음(max_depth=None), 최소 분할 샘플 수 2로 설정하였다. max_depth=None은 각 트리가 완전히 성장하도록 허용하는 설정으로, 단일 트리에서는

과적합 위험이 있으나, 랜덤 포레스트는 다수 트리의 앙상블을 통해 분산을 감소시키므로 이 위험이 완화된다(Breiman, 2001). 검증 셋에서 max_depth={10, 20, 50, None}을 비교한 결과, 깊이 제한 없는 모델이 AUC-ROC 0.91로 가장 높은 성능을 보였으며, max_depth=20에서는 0.88, max_depth=10에서는 0.84로 나타나 완전 성장이 최적임을 확인하였다.

그래디언트 부스팅(Gradient Boosting)은 순차적 앙상블 학습을 통해 예측 성능을 높이는 모델로, 100개 추정기, 학습률 0.1, 최대 깊이 3으로 설정하였다. 세 모델의 하이퍼파라미터 설정과 선택 근거를 Table 3에 정리하였다.

Table 3. Hyperparameter Settings and Selection Rationale

Model	Parameter	Value	Rationale
Logistic Regression	C	1.0	Default; AUC difference <0.01 across C={0.01,0.1,1.0,10.0}
	max_iter	1000	Ensures convergence
	class_weight	balanced	Class imbalance correction
Random Forest	n_estimators	100	Performance converges after 100
	max_depth	None	Full growth; best AUC among depth={10,20,50,None}
	min_samples_split	2	Default; overfitting mitigated by ensemble
	class_weight	balanced	Class imbalance correction
Gradient Boosting	n_estimators	100	Converges at learning_rate=0.1
	learning_rate	0.1	Default; stable learning
	max_depth	3	Default; weak learner principle for boosting

Note: Hyperparameter search was conducted on the validation set. The test set was used only for final evaluation.

2) 클래스 불균형 처리

심한 클래스 불균형(0.21%)을 처리하기 위해 세 가지 방법을 비교 적용하였다. 랜덤 언더샘플링(Random Undersampling)은 다수 클래스를 무작위로 제거하여 소수 클래스와

다수 클래스의 비율을 1:10으로 조정하는 방법이다. SMOTE(Synthetic Minority Over-sampling Technique)는 소수 클래스 샘플 간 k-최근접 이웃을 활용하여 합성 샘플을 생성하는 방법으로, k=5를 적용하였다. SMOTE와, 마찬가지로 클래스 불균형 처리 기법인 Tomek Links의 결합은, SMOTE로 오버샘플링한 후 Tomek Links로 경계 샘플을 제거하여 결정 경계를 명확히 하는 방법이다.

3) 성향점수매칭(PSM) 인과분석

언어적 특성의 인과효과를 추정하기 위해 성향점수매칭(Propensity Score Matching; PSM)을 적용하였다. PSM은 관찰 데이터에서 실험적 상황을 모방하여 처치효과를 추정하는 방법으로, 처치집단과 통제집단 간 공변량 분포를 균형화함으로써 선택편의(selection bias)를 통제한다(Rosenbaum & Rubin, 1983).

연구에서는 각 언어적 특성 변수(분노 점수, 구체성 점수, 텍스트 길이, 감성 점수)에 대해 다음 절차로 분석을 수행하였다. 먼저 처치 정의 단계에서 해당 변수의 중앙값을 기준으로 처치집단(중앙값 초과)과 통제집단(중앙값 이하)을 구분하였다. 이어서 성향점수 추정 단계에서 로지스틱 회귀를 통해 처치 할당 확률(성향점수)을 추정하였으며, 공변량으로는 처치변수를 제외한 나머지 특성 변수를 투입하였다. 매칭 단계에서는 최근접 이웃 매칭(Nearest Neighbor Matching)으로 처치-통제 쌍을 구성하였으며, caliper=0.1을 적용하였다. 마지막으로 효과 추정 단계에서 매칭된 표본에서 처치집단 평균 처치효과(Average Treatment Effect on the Treated; ATT)를 산출하였다. ATT는 다음과 같이 정의된다. 여기서 Y_1은 처치 시 잠재결과, Y_0은 비처치 시 잠재결과, T는 처치 여부를 나타낸다.

$$ATT = E[Y_1 - Y_0 \mid T = 1] = \bar{Y}\_treated - \bar{Y}\_control$$

PSM 처치 정의의 이론적 근거로서, 중앙값을 기준으로 한 이분화는 다음 논리에 근거한다. 첫째, 상대적 비교 관점에서 "다른 청원보다 더 긴/강한/구체적인" 청원의 효과를 추정한다. 둘째, 매칭을 통해 다른 조건이 동일한 청원 간 비교가 가능해진다. 셋째, 연속변수를 그대로 활용하는 일반화 성향점수(GPS) 방법은 해석의 직관성이 낮아 이분화를 선택하였다.

연구의 주요 변수(분노 점수, 구체성 점수 등)는 우편향(right-skewed) 분포를 보인다(Table 5a에서 평균이 중앙값보다 큰 것이 이를 반영함). 우편향 분포에서 평균값을 기준으로 이분화하면 처치집단의 비율이 50% 미만으로 축소되어 처치-통제 간 비대칭이 발생하고, 극단값의 영향을 받는다. 반면, 중앙값은 분포의 형태에 영향을 받지 않아(robust) 항상 약 50:50의 균등 분할을 보장하며, 이는 PSM에서 충분한 매칭 쌍을 확보하는 데 유리하다. 실제로 평균값 기준으로 재분석한 결과, 효과의 방향과 통계적 유의성은 동일하였으나 매칭률이 5~15%p 감소하여 중앙값 기준의 우월성을 확인하였다.

매칭 품질을 검증하기 위해 다음 진단을 수행하였다. 공변량 균형 검정에서는 매칭 전후 처치집단과 통제집단 간 공변량 분포의 균형 정도를 표준화 평균 차이(Standardized Mean Difference; SMD)로 평가하였으며, SMD < 0.1을 양호한 균형의 기준으로 삼았다. 공통지지 영역(Common Support) 검토에서는 성향점수 분포의 중첩 정도를 확인하여, 극단적 성향점수를 가진 관측치가 매칭에서 제외되도록 caliper=0.1을 적용하였다.

Table 4. Covariate Balance Test Results (Treatment: text_length)

Covariate	Pre-matching SMD	Post-matching SMD
word_count	0.847	0.043
sentence_count	0.312	0.028
anger_score	0.156	0.067
specificity_score	0.289	0.051
sentiment_score	0.087	0.032

Note: All post-matching SMDs fell below 0.1, confirming satisfactory balance.

4) 상호작용 효과 분석

언어적 특성 간 상호작용 효과를 검증하기 위해 조건별 정책 의제화 성공률을 비교하였다. 각 특성(분노, 구체성, 길이)을 중앙값 기준으로 고/저로 분류하고, 8가지 조합(2^3)의 성공률을

분석하였다. 시너지 효과는 다음과 같이 정의하였다. 여기서 A는 분노 고/저, S는 구체성 고/저를 나타낸다.

$$\text{Synergy} = P(Y=1|A=1,S=1) - [P(Y=1|A=1,S=0) + P(Y=1|A=0,S=1) - P(Y=1|A=0,S=0)]$$

5) 평가 지표

분류 모델의 성능은 다양한 지표로 평가하였다. 정확도(Accuracy)는 전체 예측 중 정답 비율을 나타내며, 정밀도(Precision)는 양성 예측 중 실제 양성 비율을, 재현율(Recall)은 실제 양성 중 양성 예측 비율을 측정한다. F1-Score는 정밀도와 재현율의 조화평균으로 두 지표 간 균형을 반영한다. AUC-ROC는 ROC 곡선 아래 면적으로 분류기의 전반적 성능을 측정하며, PR-AUC(Average Precision)는 정밀도-재현율 곡선 아래 면적으로 심한 불균형에서 AUC-ROC보다 민감한 지표이다. MCC(Matthews Correlation Coefficient)는 불균형 데이터에서 신뢰할 수 있는 상관계수로 활용된다. 클래스 불균형 상황에서는 정확도보다 AUC-ROC, PR-AUC, MCC가 더 적절한 평가 지표이므로, 모델 비교의 주요 기준으로 이들을 활용하였다.[1)]

IV. 분석 결과

1. 기술통계

1) 분석은 Python 3.12 환경에서 수행하였다. 주요 라이브러리로는 pandas 2.1.3, numpy 1.26.2, scikit-learn 1.3.2, scipy 1.11.4, imbalanced-learn 0.11.0을 사용하였다. 재현성 확보를 위해 모든 분석에서 랜덤 시드(random seed=42)를 고정하였다. 전체 분석 코드는 GitHub 리포지토리(https://github.com/LeeSeogMin/korean-petition.git)를 참조바란다.

Table 5a. Descriptive Statistics of Key Variables

Variable	Mean	SD	Min	Median	Max
Text length (chars)	423	335	50	331	8,050
Word count	94	75	10	74	1,949
Sentence count	8.5	6.8	1	7	150
Sentiment score	−0.047	0.231	−1	−0.042	1
Anger score	0.042	0.067	0	0.021	1
Specificity score	0.032	0.032	0	0.022	0.39

연구에 사용한 청원 텍스트의 평균 길이는 423자, 평균 단어 수는 94개였다. 감성 점수의 평균은 −0.047로 전반적으로 부정적 감성이 우세하였다. 분노 점수와 구체성 점수의 평균은 각각 0.042, 0.032로 나타났다. 정책 의제화(10,000명 이상 동의)에 성공한 청원은 전체의 0.21%(898건)에 불과하여 극심한 클래스 불균형이 확인되었다.

Table 5b. Comparison of Key Variables by Policy Agenda-Setting Status

Variable	Non-success (n=427,005)	Success (n=898)	Difference	Cohen's *d*
Text length (chars)	422	810	+388	1.16
Word count	94	181	+87	1.16
Anger score	0.042	0.059	+0.017	0.25
Specificity score	0.032	0.049	+0.017	0.53
Sentiment score	−0.047	−0.054	−0.007	0.03

Note: Cohen's dinterpretation: small (0.2), medium (0.5), large (0.8).

Table 5b에서 정책의제화 성공 청원은 미성공 청원 대비텍스트 길이가 약 1.9배(Cohen's d=1.16, 대효과), 구체성 점수가 약 1.5배(Cohen's d=0.53, 중효과) 높았다. 분노 점수도 성공 청원에서 다소 높았으나 효과 크기는 소효과 수준(Cohen's d=0.25)이었다. 감성 점수는

실질적 차이가 없었다(Cohen's d=0.03). 이는 표본 수가 매우 크기 때문에 p-value가 극히 작게 나타날 수 있으므로, 효과 크기를 함께 고려하여 실질적 유의성을 판단해야 함을 시사한다.

2. 머신러닝 분류 모델 성능

1) 클래스 불균형 처리 방법 비교

Table 6. Performance by Class Imbalance Method (Random Forest, Test Set)

Method	Accuracy	Precision	Recall	F1	AUC-ROC	PR-AUC	MCC
Undersampling	0.986	0.047	0.289	0.080	0.910	0.043	0.112
SMOTE	0.988	0.038	0.185	0.063	0.881	0.029	0.080
SMOTE+Tomek	0.989	0.041	0.200	0.069	0.869	0.028	0.087

Note: Bold values indicate best performance per metric. AUC-ROC, PR-AUC, and MCC (bold column headers) are primary evaluation criteria for imbalanced data.

Table 6을 보면, 언더샘플링이 SMOTE 기반 방법보다 우수한 성능을 보였다(AUC-ROC 0.91 vs 0.88). 이는 심한 불균형(0.21%) 상황에서 SMOTE가 생성하는 합성 샘플이 노이즈를 유발할 수 있음을 시사한다. 따라서 본 연구에서는 언더샘플링을 주요 분석 방법으로 채택하였다.

2) 모델별 성능 비교

Table 7. Baseline Model Performance (Undersampling, Test Set)

Model	Accuracy	Precision	Recall	F1	AUC-ROC	PR-AUC	MCC
Logistic Regression	0.784	0.007	0.669	0.013	0.785	0.012	0.041
Random Forest	0.986	0.047	0.289	0.080	0.910	0.043	0.112
Gradient Boosting	0.983	0.041	0.324	0.073	0.899	0.038	0.087

Note: Bold values indicate best performance per metric. AUC-ROC, PR-AUC, and MCC (bold column headers) are primary evaluation criteria.

Table 7에서 Random Forest 모델이 AUC-ROC 0.910, PR-AUC 0.043, MCC 0.112로 세 가지 핵심 지표 모두에서 가장 높은 성능을 보였다. 이는 언어적 특성만으로도 정책의제화 여부를 상당 수준 변별할 수 있음을 시사한다. 다만, 극심한 클래스 불균형(양성 비율 0.21%)으로 인해 모든 모델의 F1-Score(0.013~0.080)와 재현율(0.185~0.669)이 낮은 수준에 머물렀다. 이는 해당 모델들이 현재 상태로는 정책의제화 조기 경보 시스템으로 직접 활용하기에는 한계가 있음을 의미한다. 특히 Random Forest의 재현율이 0.289에 그쳐, 실제 성공 청원 10건 중 약 3건만을 포착한다. Logistic Regression은 높은 재현율(0.669)을 보이나 정밀도(0.007)가 극히 낮아 위양성 비율이 과도하다. 따라서 본 연구의 머신러닝 분석은 실용적 예측 도구보다는, 언어적 특성이 정책의제화에 기여하는 상대적 중요도를 파악하는 탐색적 도구로 해석하는 것이 적절하다. 향후 Focal Loss, 비용 민감 학습(cost-sensitive learning) 등 고급 불균형 처리 기법과 BERT 기반 임베딩을 결합하면 예측 성능의 실질적 향상을 기대할 수 있다.

3) 변수 중요도 분석

Table 8. Random Forest Variable Importance

Rank	Variable	Importance	Permutation Importance
1	word_count	0.273	0.0002
2	text_length	0.251	0.0002
3	sentence_count	0.168	0.0004
4	anger_score	0.119	0.0003
5	specificity_score	0.117	−0.0004
6	sentiment_score	0.072	−0.0001

Table 8에서 단어 수와 텍스트 길이가 가장 중요한 예측 변수로 나타났으며, 분노 점수와 구체성 점수가 그 다음 순위를 차지하였다. 감성 점수(긍정/부정)는 상대적으로 낮은 중요도를 보였다. 통제변수(category_encoded, topic_id, year, month)는 머신러닝 예측 모델에 포함하

지 않았으므로, 변수 중요도 분석에서 제외되었다. avg_sentence_length와avg_word_length는 기존 변수(text_length, sentence_count, word_count)의 파생 지표로서, 모델에 포함하였으나 기여도가 미미하여(각각 중요도 < 0.001) 표에서 생략하였다.

3. PSM 인과분석 결과

Table 9. PSM Results (ATT Estimates)

Treatment Variable	ATT (%p)	SE	t-statistic	p-value	Matched Pairs	Match Rate
Text length	0.312***	0.0002	18.99	<0.001	149,607	69.9%
Anger score	0.212***	0.0004	5.45	<0.001	50,932	23.8%
Specificity score	0.126***	0.0002	6.81	<0.001	149,760	70.0%
Sentiment score	−0.042	0.0002	−1.70	0.090	69,550	32.5%

Note: *** p<0.001, ** p<0.01, * p<0.05. Match rate = proportion of treated units successfully matched.

Table 9에서 PSM 분석 결과, 텍스트 길이, 분노 점수, 구체성 점수가 정책의제화에 유의한 인과적 영향을 미치는 것으로 나타났다. 텍스트 길이는 가장 강한 효과를 보였으며(ATT=0.312%p, p<0.001), 기본 비율(0.21%)을 고려하면 약 148%의 상대적 증가에 해당한다. 분노 점수도 유의한 양(+)의 효과를 보였으며(ATT=0.212%p, p<0.001), 분노 표현이 강한 청원이 더 높은 정책의제화 확률을 보인다는 점은 선행연구(Song & Park, 2019; Hagen et al., 2016)의 발견과 일치한다. 구체성 점수 역시 유의한 양(+)의 효과를 나타냈다. 반면 감성 점수(긍정/부정)는 유의한 효과를 보이지 않았다. 이는 감정의 방향 자체보다는 감정의 강도(특히 분노)가 정책의제화에 더 중요함을 시사한다.

4. 상호작용 효과 분석

언어적 특성 간 상호작용 효과를 검증하기 위해 교호항 로지스틱 회귀를 적합하였다(Table 10).

Table 10. Interaction Term Logistic Regression Results

Variable	OR	95% CI	z	p-value
Main Effects				
anger_high	1.464	[1.393, 1.540]	14.91	<0.001***
specificity_high	1.174	[1.133, 1.217]	8.82	<0.001***
length_high	2.698	[2.615, 2.783]	62.18	<0.001***
2-way Interactions				
anger_high × specificity_high	2.126	[1.958, 2.308]	17.97	<0.001***
anger_high × length_high	1.605	[1.502, 1.714]	14.07	<0.001***
specificity_high × length_high	1.678	[1.605, 1.755]	22.78	<0.001***
3-way Interaction				
anger_high ×specificity_high × length_high	0.563	[0.512, 0.620]	−11.73	<0.001***

Note: *** p<0.001. Controls (category, year, month) included but not shown. OR = Odds Ratio. class_weight='balanced' applied.

주효과 분석 결과, 텍스트 길이가 가장 강한 효과를 보여(OR=2.698, p<0.001), 중앙값 이상의 길이를 가진 청원이 그렇지 않은 청원 대비 정책의제화 승산이 약 2.7배 높았다. 분노 표현(OR=1.464, p<0.001)과 구체성(OR=1.174, p<0.001)도 유의한 주효과를 보였다.

2원 교호항은 모두 통계적으로 유의하였다. 분노*구체성 교호항(OR=2.126, p<0.001)이 가장 강한 시너지 효과를 보여, 두 특성이 동시에 높을 때 개별 효과의 합산을 넘어서는 추가적 승산 증가가 발생하였다. 분노*길이(OR=1.605, p<0.001)와 구체성*길이(OR=1.678, p<0.001) 교호항도 유의한 양의 시너지를 나타냈다.

3원 교호항(분노*구체성*길이)은 유의하나 OR이 1 미만(OR=0.563, p<0.001)으로 나타났다. 이는 세 특성이 동시에 높을 때 2원 교호항들의 시너지가 다소 감쇄됨을 의미한다. 즉, 세 특성 모두 높은 조건에서의 효과는 세 개의 2원 교호항이 시사하는 것보다 약화되나, 주효과와 2원 교호항의 강한 양(+)의 효과로 인해 전체적인 정책의제화 승산은 여전히

크게 증가한다.

5. 임계값 민감도 분석

Table 11. Sensitivity Analysis: PSM Results Across Thresholds

Threshold	Positive Rate	Positive N	Text Length ATT	Anger Score ATT	Specificity ATT
5,000	0.42%	1,797	0.28%p***	0.19%p***	0.11%p***
10,000	0.21%	898	0.31%p***	0.21%p*	0.13%p*
20,000	0.11%	471	0.34%p***	0.23%p**	0.14%p**
50,000	0.04%	189	–	–	–
100,000	0.02%	98	–	–	–
200,000	0.015%	65	–	–	–

Note: *** p<0.001, ** p<0.01. Thresholds ≥50,000 used t-tests instead of PSM due to sample size constraints.

결과를 보면, 다양한 임계값에서 주요 변수들의 효과 방향과 통계적 유의성이 일관되게 유지되어, 본 연구 결과의 강건성을 확인하였다. 청와대 공식 답변 기준인 200,000명 이상 동의를 받은 65건의 청원에 대해 추가 분석을 수행하였다. 표본 수 제약으로 PSM은 적용할 수 없었으나, t-검정 결과 10,000명 기준과 동일한 방향의 효과를 확인하였다.

Table 12. Comparison at Official Threshold (200,000 Signatures)

Variable	200k+ Petitions (n=65)	Other Petitions	Difference	t-statistic	p-value
Text length (chars)	1,113	422	+691	14.13	<0.001
Anger score	0.074	0.035	+0.039	7.65	<0.001
Specificity score	0.065	0.032	+0.033	7.64	<0.001

200,000명 이상 동의를 받은 청원은 평균 텍스트 길이가 2.6배, 분노 점수가 2.1배, 구체성 점수가 2.0배 높았으며, 모든 차이가 통계적으로 유의하였다($p<0.001$). 이는 본 연구의 주요 발견이 공식 기준에서도 일관되게 유지됨을 의미한다.

6. 가설 검증 결과

Table 13. Hypothesis Test Results

Hypothesis	Content	Result	Evidence
H1	Petitions with stronger anger expressions have higher policy agenda probability	Supported	ATT=0.21%p, $p<0.001$
H2	Petitions with specific information have higher policy agenda probability	Supported	ATT=0.13%p, $p<0.001$
H3	Longer petitions have higher policy agenda probability	Supported	ATT=0.31%p, $p<0.001$
H4	Synergy effect occurs when anger, specificity, and length are simultaneously high	Supported	All 2-way interaction ORs >1 ($p<0.001$); anger×specificity OR=2.126

Table 13을 보면, 네 가지 연구 가설이 모두 지지되었음을 확인할 수 있다. 특히 상호작용 효과의 발견은 개별 특성의 효과를 넘어서는 새로운 통찰을 제공한다.

V. 논의 및 결론

1. 학문적 함의

전통적으로 의제 설정 과정은 미디어, 정치인, 이익집단의 역할에 초점을 맞추어왔으나 (Kingdon, 1984), 이 연구는 디지털 플랫폼에서 개별 시민의 메시지 구성 방식이 의제화에

직접적 영향을 미침을 실증하였다.

특히 상호작용 효과의 발견은 이론적으로 중요한 함의를 갖는다. 분노(감정적 요소)와 구체성(인지적 요소)의 시너지 효과는 감정과 이성이 상호보완적으로 작용하여 시민 동원에 기여함을 시사한다. 이는 듀얼 프로세스 이론(dual-process theory)과 일관된 발견이다.

방법론적으로는 대규모 한국어 텍스트 데이터에 머신러닝 예측과 PSM 인과분석을 결합한 분석 프레임워크를 제시하였다. 또한 극심한 클래스 불균형(0.21%) 상황에서 언더샘플링이 SMOTE보다 효과적임을 실증하였으며, 상호작용 효과 분석을 통한 복합적 인과 메커니즘을 규명하였다.

2. 정책적 함의

시민 관점에서 성공적인 청원을 위해서는 충분한 길이로 문제를 상세히 설명하는 것이 가장 중요하며, 구체적인 사실과 근거(숫자, 날짜, 사례 등)를 제시하고 적절한 감정적 호소(특히 분노, 억울함)를 포함하는 것이 효과적이다. 특히 이 세 요소를 동시에 갖출 때 시너지 효과가 발생하여 정책의제화 승산이 크게 증가한다(분노×구체성 OR=2.126, $p<0.001$).

정부 관점에서는 청원의 양적 지표(동의 수)만이 아니라 질적 지표(논리성, 구체성)를 함께 고려해야 한다. 본 연구의 분석 프레임워크는 언어적 특성의 상대적 중요도를 파악하는 데 활용할 수 있으며, 향후 딥러닝 기반 모델과 결합하여 잠재적 주요 청원 조기 식별 시스템으로 발전시킬 수 있다.

3. 한계 및 향후 연구

감성 분석 방법의 측면에서, 본 연구에서는 사전 기반 방법을 사용하여 감성과 분노를 측정하였다. 이 방법은 한국어의 맥락적 의미, 부정(否定) 표현, 비유적 표현 등을 충분히 반영하지 못할 수 있다. 향후 연구에서는 KoBERT, KoELECTRA 등 사전훈련 언어모델을 활용한 정교한 감성 분석을 적용하여 측정 타당성을 높일 필요가 있다.

PSM의 방법론적 측면에서, PSM은 관찰되지 않은 혼란변수(unobserved confounders)에 대해 취약하다. 예를 들어, 청원 작성자의 사회적 네트워크, 미디어 보도 여부, 정치적 시의성

등 관찰되지 않은 요인이 결과에 영향을 미쳤을 가능성이 있다. 또한 연속변수를 중앙값 기준으로 이분화한 것은 정보 손실을 야기할 수 있다. 향후 연구에서는 일반화 성향점수(GPS), 도구변수(IV) 분석, 이중차분법(Difference-in-Differences) 등 대안적 인과추론 방법의 적용을 검토할 필요가 있다.

시계열적 요소와 외부 요인의 통제 측면에서, 청원 등록 시점의 정치적 상황, 미디어 보도, 사회적 이슈 등 외부 요인이 정책의제화에 미치는 영향을 충분히 통제하지 못하였다. 또한 2019년 8월 이후의 데이터를 포함하지 않아, 시간적 일반화에 제약이 있다.

향후 연구에서는 KoBERT 기반 딥러닝 분류, 시계열 분석, 미디어 데이터와의 융합 분석 등을 통해 본 연구의 한계를 보완할 수 있을 것이다.

References

Berger, J., & Milkman, K. L. (2012). What makes online content viral? Journal of Marketing Research, 49(2), 192-205.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

Chen, W., Wang, Y., & Huang, J. (2023). Understanding how the expression of online citizen petitions influences government responses in China. Information Processing & Management, 60(3), 103357.

Cobb, R. W., & Elder, C. D. (1972). Participation in American Politics: The Dynamics of Agenda-Building. Johns Hopkins University Press.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT (pp. 4171-4186).

Eom, S. J. (2019). Current status and challenges of the Blue House National Petitions: Based on big data analysis results. Proceedings of the Korean Association for Public Administration Winter Conference, 2019, 1–25.

Hagen, L., Harrison, T. M., Uzuner, Ö., May, W., Fake, T., & Katragadda, S. (2016). E-petition popularity: Do linguistic and semantic factors matter? Government Information Quarterly, 33(4), 783–795.

Kingdon, J. W. (1984). Agendas, Alternatives, and Public Policies. Little, Brown.

Lovit. (2019). Blue House National Petition Archive [Data set]. GitHub. https://github.com/lovit/petitions_archive

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41–55.

SKTBrain. (2019). KoBERT: Korean BERT pre-trained cased. GitHub. https://github.com/SKTBrain/KoBERT

Song, J. M., & Park, Y. D. (2019). What happens in the Blue House National Petitions?: An analysis of the Blue House National Petitions using natural language processing. Korean Political Science Review, 53(5), 53–78.

Wright, S. (2016). 'Success' and online political participation: The case of Downing Street E-petitions. Information, Communication & Society, 19(6), 843–857.

이석민(Lee, Seog-Min) : 서울대학교에서 인류학(학부), 의회정치(석사), 과학기술정책(박사)을 수학했으며, 박사후 과정에서 통계학과 계량경제학을 연구했다. 인류학적 관찰에서 시작하여 정치학과 정책학의 이론적 토대를 거쳐, 계량경제학과 통계학의 방법론적 엄밀성을 습득한 후, 최근에는 딥러닝과 생성형 AI를 정책 현장에 적용하는 연구를 수행하고 있다. 수원대 학교 행정학과에서 빅데이터 분석을 연구했으며, 현재 한신대학교 공공인재빅데이터융 합학부에서 AI 기반 정책분석과 딥러닝 응용 연구를 진행하고 있다. 『AI를 활용한 시니어 정신건강 진단 앱』을 개발하고 특허출원했다. 저서로는 『R과 STATA를 활용한 정책평가방법론』(2017), 『빅데이터분석방법론』, 『구조방정식: 준실험설계접근』, 『인과 데이터 사이언스 (Causal Data Science): AI 시대를 위한 분석·평가 모델링의 진화』 (2026), 『LLM 시대의 인문사회과학방법론』 (2026)등이 있으며, 행정학 및 이공계 학술지에 논문을 게재하였다. 웹프로그래밍, 데이터베이스, MLOps, 딥러닝 자연어 처리 및 영상 처리, AI 에이전트, 정책분석평가, 분석기획론 등 인문사회과학과 AI·Software 기술을 융합하는 교육을 실천하고 있다. (e-mail: newmind68@hs.ac.kr)

<논문접수일: 2025-12-31 / 논문수정일: 2026-02-14 / 게재확정일: 2026-02-14>

국민청원 텍스트의 언어적 특성이 정책의제화에 미치는 영향:

머신러닝 기반 예측 및 인과분석

이석민*

<초록>

본 연구는 청와대 국민청원 플랫폼(2017-2019)에 접수된 427,903건의 청원을 분석하여 텍스트의 언어적 특성이 정책의제화에 미치는 영향을 실증적으로 검토하였다. 텍스트 길이, 분노 표현, 구체성 등 12개 언어적 특성을 추출하고, 머신러닝 분류 모델과 성향점수매칭(PSM)을 결합하여 예측 성능과 인과효과를 분석하였다. 랜덤 포레스트 모델은 AUC-ROC 0.91을 달성하였으며, PSM 분석 결과 텍스트 길이(ATT, +0.31%p), 분노 점수(ATT, +0.21%p), 구체성 점수(ATT, +0.13%p)가 정책의제화에 유의한 인과효과를 보인 반면, 감성 극성은 유의하지 않았다. 교호항 로지스틱 회귀 분석 결과, 분노*구체성(OR=2.126), 분노*길이(OR=1.605), 구체성*길이(OR=1.678)의 2원 교호항이 모두 유의한 시너지 효과를 보였다. 이러한 결과는 정책의제화가 감정의 방향이 아닌 표현의 강도와 구체성에 의해 결정됨을 시사한다.

주제어: 국민청원, 정책의제화, 텍스트 마이닝, 머신러닝, 성향점수매칭, 감성분석

* 한신대 공공인재빅데이터융합학