

## 📄 Day 5 완료 보고서: TF-IDF 최적화 실험

### 🎯 Day 5 목표 및 달성도

#### 설정 목표

- 1차 목표: 79.40% → 80%+ 정확도 달성
- 2차 목표: TF-IDF 하이퍼파라미터 체계적 최적화
- 3차 목표: Phase 3 딥러닝 진입을 위한 탄탄한 기반 구축

#### 최종 달성 결과

목표	달성도	결과
80% 정확도	✅ 초과 달성	80.77% 달성 (+0.77%p 초과)
TF-IDF 최적화	✅ 완료	4가지 실험 체계적 수행
기반 구축	✅ 완료	Phase 3 준비 완료

### 📊 핵심 성과 요약

#### 🏆 성능 개선 결과

베이스라인 앙상블: 79.41%  
최적화 앙상블: 80.77%  
개선 효과: +1.36%p  
목표 초과: +0.77%p

#### ⚙️ 최적 설정 발견

- N-gram 범위: (1, 1) - 단일 단어
- 최대 특성 수: 20,000개 (기존 10,000개에서 2배 증가)
- 문서 빈도: min\_df=1, max\_df=1.0 (필터링 없음)
- 모델: 나이브 베이즈 + 로지스틱 회귀 앙상블

### 🔬 실험 과정 및 결과

#### 실험 1: N-gram 범위 최적화

실험 설계: 단어 조합 범위가 성능에 미치는 영향 분석

N-gram 범위	정확도	CV 평균	특성 수	결과
(1, 1)	79.41%	79.45%	10,000	🏆 최고
(1, 2)	79.15%	79.24%	10,000	-0.26%p
(1, 3)	79.06%	79.20%	10,000	-0.35%p

핵심 발견:

- 단일 단어가 최적: 한국어 리뷰의 직접적 표현 특성
- 2-gram/3-gram 효과 없음: 짧은 리뷰(평균 35.2자)에서는 단어 조합보다 핵심 단어 존재가 중요

실험 2: 특성 수 최적화 ★ 가장 큰 효과

실험 설계: 어휘 크기가 모델 성능에 미치는 영향

특성 수	정확도	CV 평균	학습 시간	개선 효과
5,000개	77.83%	77.77%	0.51초	-1.58%p
10,000개	79.41%	79.45%	0.56초	베이스라인
15,000개	80.20%	80.15%	0.58초	+0.79%p
20,000개	80.62%	80.57%	0.60초	+1.21%p

핵심 발견:

- 특성 수 증가 = 성능 향상: 더 많은 어휘로 미묘한 감정 표현 포착
- 희귀 단어의 중요성: "여운이", "수작" 같은 저빈도 감정 표현어 활용
- 계산 비용 대비 효과: 2배 특성 증가로 1.21%p 향상, 학습 시간은 7% 증가

실험 3: 문서 빈도 조정

실험 설계: 너무 희귀하거나 흔한 단어 필터링 효과

설정	min_df	max_df	정확도	실제 특성	결과
baseline	1	1.0	80.62%	20,000	🏆 최고
light_filter	2	0.95	80.58%	20,000	-0.04%p
medium_filter	3	0.90	80.60%	20,000	-0.02%p
strong_filter	5	0.85	80.59%	19,878	-0.03%p

핵심 발견:

- 필터링 효과 없음: 모든 필터링이 오히려 성능 저하
- 원본 데이터 활용: NSMC의 고품질 데이터에서는 추가 필터링 불필요
- 희귀 단어도 유용: 1회만 등장하는 단어도 감정 신호로 활용 가능

## 실험 4: 최종 모델 구축

최적 설정 적용 결과:

- 검증 데이터: 80.62%
- 테스트 데이터: 80.77% (최종 성능)
- 일반화 성능: 검증보다 테스트에서 0.15%p 높음 (과적합 없음)

## 💡 프로젝트 인사이트

### 1. NSMC 데이터 특성 재확인

텍스트 특징:

- 평균 길이: 35.2자 (극도로 간결)
- 핵심 단어: "최고"(93.2% 긍정), "최악"(98.8% 부정)
- 이모티콘: "ㅎㅎ"(84.2% 긍정), "???"(70.3% 부정)

모델링 시사점:

- 단순한 접근이 효과적: 복잡한 N-gram보다 단일 단어
- 어휘 다양성 중요: 20,000개 특성으로 미묘한 표현 포착
- 직접적 감정 표현: 한국어 리뷰의 솔직하고 간결한 특성

### 2. TF-IDF 최적화 전략

성공 요인:

- 체계적 실험: 한 번에 하나씩 파라미터 조정
- 데이터 기반 선택: 이론적 가정보다 실제 성능 우선
- 계산 비용 고려: 성능 향상 대비 합리적 자원 사용

실패 요인:

- 이론적 편견: N-gram이 더 좋을 것이라는 예상
- 과도한 필터링: 좋은 데이터에 불필요한 전처리

### 3. 한국어 NLP 특성

언어적 특성:

- 응축적 표현: 짧은 문장에 강한 감정 표현
- 직접적 소통: 영어권 대비 더 직설적인 평가
- 이모티콘 활용: 감정 증폭 수단으로 적극 활용

## 모델링 영향:

- 복잡한 구문 분석보다 **단어 단위 감정 점수**가 효과적
- **문화적 맥락**을 반영한 감정 어휘 중요성

## 기술적 성취

### 구현 완료 사항

#### 1. 체계적 실험 프레임워크

- 4가지 하이퍼파라미터 실험 설계
- 교차 검증 기반 안정적 성능 측정
- 자동화된 결과 비교 및 시각화

#### 2. 최적화된 파이프라인

```
python

# 최종 최적 설정
TfidfVectorizer(
    max_features=20000, # 2배 증가
    ngram_range=(1, 1), # 단일 단어
    min_df=1, # 필터링 없음
    max_df=1.0
)

# 앙상블 모델
VotingClassifier([
    ('nb', MultinomialNB(alpha=1.0)),
    ('lr', LogisticRegression(C=1.0))
], voting='soft')
```

#### 3. 성능 모니터링

- 실시간 성능 추적
- 계산 비용 대비 효과 분석
- 시각화 기반 결과 해석

## 생성된 자산

### 모델 파일:

- `optimized_ensemble_v1.pkl`: 최적화된 앙상블 모델

- `optimized_tfidf_v1.pkl`: 최적화된 TF-IDF 벡터라이저

### 분석 결과:









- `tfidf_optimization_results.png`: 실험 결과 시각화
- `bow_tfidf_optimization.ipynb`: 재현 가능한 실험 노트북

### 성능 기록:

- 체계적인 실험 로그 및 성능 지표
- 하이퍼파라미터별 상세 분석 결과

## 프로젝트 진행도


### 현재까지 달성한 마일스톤

- Phase 1: 데이터 이해 (Day 1-3) 
- 환경 설정 및 프로젝트 구조 
  - NSMC 데이터 탐색 및 특성 분석 
  - 한국어 텍스트 특성 발견 
- Phase 2: 기본 모델 개발 (Day 4-5) 
- 머신러닝 모델 비교 (79.40%) 
  - TF-IDF 최적화 (80.77%) 
  - 80% 목표 달성 

### 성능 발전 궤적

단계	모델	정확도	누적 개선
Day 4	나이브 베이즈	79.25%	+4.25%p
Day 4	로지스틱 회귀	79.17%	+4.17%p
Day 4	앙상블	79.40%	+4.40%p
Day 5	최적화 앙상블	80.77%	+5.77%p

### 목표 대비 달성률

- 초기 계획 목표: 75% → 80% → 85% → 90%
- 실제 달성 현황: 80.77%  (1단계 목표 초과 달성)
- 다음 목표: 85%+ (딥러닝)
- 최종 목표: 90%+ (Transformer)

## 🎯 다음 단계 전략

### Phase 3: 딥러닝 모델 개발 (예정)

목표: 80.77% → 85%+ 달성

예정 모델들:

1. LSTM: 시퀀스 정보 활용
2. CNN: 지역적 패턴 포착
3. Word2Vec + 딥러닝: 의미론적 임베딩

기대 효과:

- 순서 정보: "정말 재미있어요"의 단어 순서 활용
- 문맥 이해: "기분 좋게 틀었는데 잡혔다" 같은 복합 표현
- 의미 관계: 유사 단어 간 관계 학습

### 활용할 최적화 경험

Day 5에서 얻은 교훈 적용:

1. 데이터 중심 접근: 이론보다 실험 결과 우선
2. 체계적 실험: 한 번에 하나씩 변수 조정
3. 한국어 특성 반영: 직접적 표현의 중요성

---

## ✅ Day 5 완료 체크리스트

### 🎯 목표 달성도

- ✅ 80% 정확도 달성: 80.77% (목표 초과)
- ✅ TF-IDF 최적화: 4가지 실험 완료
- ✅ 체계적 문서화: 재현 가능한 실험 기록
- ✅ 모델 저장: 최적화된 모델 보존

### 📚 핵심 학습 성과

- ✅ 하이퍼파라미터 튜닝: 체계적 최적화 프로세스
- ✅ 한국어 NLP 이해: 언어별 특성 고려의 중요성
- ✅ 실험 설계: 과학적 방법론 적용
- ✅ 성능 분석: 계산 비용 대비 효과 평가

### 🔧 기술적 완성도

- ✅ 재현 가능한 코드: 표준화된 실험 환경

- ✓ 자동화된 평가: 교차 검증 및 시각화
- ✓ 문서화: 상세한 실험 기록 및 해석
- ✓ 버전 관리: Git을 통한 체계적 관리

## 🎉 Day 5 성과 요약

### 양적 성과

- 80% 목표 달성: 계획보다 빠른 달성 (5일차)
- 1.36%p 성능 향상: 실질적이고 의미있는 개선
- 4가지 실험 완료: 체계적이고 과학적인 접근

### 질적 성과

- 한국어 NLP 전문성: 언어 특성 기반 최적화
- 실험 방법론: 재현 가능한 과학적 프로세스
- 실무 적용성: 프로덕션 환경에서 사용 가능한 모델

### 포트폴리오 가치

- 차별화된 접근: 체계적 TF-IDF 최적화
- 명확한 결과: 구체적 성능 개선 수치
- 전문성 증명: 한국어 NLP 특화 인사이트

## 📝 최종 권장사항

### 즉시 진행 사항

- Phase 3 준비: 딥러닝 모델 개발 환경 구축
- 중간 점검: 전체 프로젝트 진행 상황 검토
- 문서 정리: Day 1-5 성과 통합 정리

### 중장기 발전 방향

- 딥러닝 도전: 85% 정확도 목표
- Transformer 적용: KoBERT로 90% 도전
- 실무 적용: 웹 데모 및 API 구축

## 📅 보고서 정보

- 작성일: 2025년 1월

- **작성자:** NSMC 감정분석 프로젝트팀
- **버전:** v1.0
- **다음 단계:** Phase 3 딥러닝 모델 개발