📋 Day 3 완료 보고서: NSMC 데이터 탐색 및 한국어 텍스트 특성 분석

🎯 Day 3 목표 및 달성도

목표	달성도	결과
NSMC 데이터 다운로드 및 로딩	100%	150,000개 훈련, 50,000개 테스트 데이터 성공 로딩
기본 데이터 구조 분석	100%	완벽한 50:50 감정 분포, 결측값 없음 확인
텍스트 길이 특성 파악	100%	평균 35.2자, 82.5%가 50자 이하의 간결한 리뷰
한국어 텍스트 특성 분석	100%	이모티콘, 감정별 단어, 특수문자 패턴 발견
첫 번째 노트북 완성	100%	data_overview.ipynb 체계적 구성 완료

📊 주요 발견사항

1. 데이터 품질 및 구조

완벽한 데이터 균형

- 훈련 데이터: 긍정 74,825개 vs 부정 75,170개 (거의 완벽한 균형)
- 테스트 데이터: 동일한 50:50 분포 유지
- 결측값, 중복값 거의 없는 고품질 데이터

데이터 규모의 적정성

- 총 200,000개 리뷰는 한국어 감정분석 학습에 충분한 규모
- 메모리 사용량 약 15MB로 일반 PC에서 처리 가능
- 클래스 불균형 문제 없어 정확도 지표 신뢰 가능

2. 텍스트 길이 특성

극도로 간결한 표현 문화

- 평균 길이: 35.2자 (단어 7.6개)
- 중앙값: 27자 (평균보다 짧음, 오른쪽 꼬리 분포)
- 82.5%가 50자 이하 (트위터 초기 제한보다도 짧음)

극단적 표현의 존재

- 최단: 1자 ("아", "잼"으로 감정 표현)
- 최장: 146자 (에세이 수준의 리뷰도 존재)
- 감정별 길이 차이: 거의 없음 (부정 35.8자 vs 긍정 34.6자)

모델링 시사점

- 짧은 시퀀스 처리에 최적화된 모델 선택 가능
- 단어 하나하나의 중요도가 높음
- 패딩 길이 100자 내외로 설정하면 94.2% 커버 가능

3. 한국어 텍스트의 독특한 특성

문자 유형별 분포

- 한글: 98.9% (거의 모든 리뷰가 한국어 포함)
- 특수문자: 68.8% (감정 표현에 특수문자 적극 활용)
- 이모티콘: 12.8% (상당한 비율의 이모티콘 사용)
- 영어: 5.9%, 숫자: 13.6% (제한적이지만 의미있는 혼용)

감정별 핵심 구분 단어

- 강한 긍정 지표: "최고"(93.2%), "좋다"(87.8%), "ㅎㅎ"(84.2%)
- 강한 부정 지표: "최악"(1.2%), "별로"(16.3%), "???"(29.7%)
- 중성적 단어: "영화"(51.5%), "너무"(52.8%), "진짜"(48.7%)

이모티콘의 예상과 다른 패턴

- "ㅠㅠ": 긍정 72.7% (슬픔보다는 감탄의 의미로 사용)
- "ㅋㅋ": 긍정 56.8% (생각보다 부정에서도 빈번 사용)
- "ㅎㅎ": 긍정 84.2% (가장 확실한 긍정 신호)

🔍 심층 분석 결과

전체 단어 빈도 TOP 5

1. "영화": 10,825회 (모든 리뷰의 중심 키워드)
2. "너무": 8,239회 (강도 표현 부사)
3. "정말": 7,791회 (긍정에서 더 빈번, 67.9%)
4. "진짜": 5,929회 (감정 중립적)
5. "영화.": 3,598회 (문장 종료 패턴)

특수 표현 사용 패턴

- "...": 16.1% 사용, 부정 59.5% (가장 많이 사용되는 특수 표현)
- "!!!": 4.4% 사용, 긍정 74.4% (강조 표현)
- "ㅋㅋ": 5.0% 사용, 긍정 56.8% (웃음 표현)

극단적 사례들

- 1자 리뷰: "아"(부정), "잼"(긍정) - 한국어의 극도 축약 가능성
- 146자 리뷰: 상세한 비평문 수준의 장문 리뷰 존재
- 이모티콘 집중 사용: "ㅋㅋㅋ 진짜 골깜..ㅋㅋ 눈 부라릴때 쓰러짐..ㅋㅋ"

🚀 모델링 전략 수립

즉시 적용 가능한 인사이트

1. 특성 추출 전략

- 감정 구분력 높은 단어들("최고", "최악", "좋다") 별도 가중치 부여
- 이모티콘 패턴을 감정 점수로 변환 ("ㅎㅎ" → +0.8, "???" → -0.7)
- 특수문자 빈도를 감정 강도 지표로 활용

2. 전처리 우선순위

- 보존: 감정 지표 단어, 의미있는 이모티콘, 감정 특수문자
- 정규화: "영화."→"영화", 반복 표현 정리
- 제거 고려: 너무 일반적인 단어들("영화", "너무", "진짜")

3. 모델 아키텍처 방향

- 짧은 시퀀스에 최적화된 모델 (LSTM, CNN, Transformer 모두 가능)
- 단어 단위보다 문자 단위 처리도 고려 (1자 리뷰 대응)
- 특수문자와 이모티콘을 별도 임베딩으로 처리

한국어 특화 접근법

1. 언어적 특성 반영

- 조사, 어미 변화를 고려한 어근 추출
- 구어체 표현의 표준화 규칙 수립
- 감정 표현 축약 패턴 학습

2. 문화적 맥락 고려

- 한국 인터넷 문화의 이모티콘 사용 패턴
- 영화 리뷰 특유의 표현 방식
- 세대별, 성별 언어 사용 차이 가능성

📈 기대 성과 및 다음 단계

Day 3 완료로 달성한 가치

1. 데이터 친숙도 확보

- NSMC 데이터에 대한 완전한 이해
- 한국어 텍스트 처리 경험 습득
- 실무적 데이터 탐색 프로세스 체득

2. 분석 기반 구축

- 정량적 분석을 통한 객관적 인사이트
- 가설 기반 접근법 대신 데이터 기반 전략 수립
- 체계적인 문서화와 재현 가능한 분석 환경

3. 모델링 방향 설정

- 85% 목표 달성을 위한 구체적 전략 수립
- 한국어 NLP의 도전과제와 기회 요소 파악
- 전처리부터 모델 선택까지 전체 파이프라인 설계 기반 마련

Day 4 진행 준비도

기술적 준비 완료

- 데이터 로딩 및 처리 파이프라인 구축
- 한국어 텍스트 분석 도구 숙련도 확보
- 체계적인 노트북 구조 및 문서화 습관 정착

분석적 준비 완료

- 전처리 우선순위 및 전략 명확화
- 특성 추출 방법론 구체화
- 모델 성능 향상 포인트 식별

전략적 준비 완료

- 현실적이고 달성 가능한 목표 설정 (85% 정확도)
- 단계별 성능 개선 로드맵 수립
- 한국어 특화 접근법 방향성 확보

💡 핵심 성공 요인

체계적 접근법의 효과

- 데이터 이해 → 가설 수립 → 검증의 순차적 진행
- 정량적 분석과 정성적 해석의 균형
- 기술적 구현과 도메인 지식의 조화

한국어 NLP 특성 반영

- 영어 중심 방법론의 한계 인식
- 한국어만의 독특한 특성 발견 및 활용
- 문화적 맥락을 고려한 텍스트 해석

실무 지향적 사고

- 학술적 완벽성보다 실용적 효과 추구
- 모델 성능과 개발 효율성의 균형
- 재현 가능하고 확장 가능한 솔루션 지향

🎯 결론 및 다음 단계

Day 3는 NSMC 프로젝트의 **탄탄한 기초**를 구축했다. 데이터의 모든 특성을 파악하고, 한국어 텍스트만의 독특한 패턴들을 발견했으며, 이를 바탕으로 85% 정확도 달성을 위한 구체적인 전략을 수립했다.

특히 "ㅎㅎ"가 84.2% 긍정 지표이고, "???"가 70.3% 부정 지표라는 발견은 한국어 감정분석에서 이모티콘의 중요성을 보여준다. 또한 평균 35.2자의 극도로 짧은 리뷰에서도 명확한 감정 표현이 가능하다는 점은 한국어의 효율적 표현력을 증명한다.

Day 4 예정 작업: 이 분석 결과를 바탕으로 한국어 특화 텍스트 전처리 파이프라인을 구축하고, 첫 번째 베이스라인 모델을 통해 75% 이상 정확도 달성을 목표로 한다.