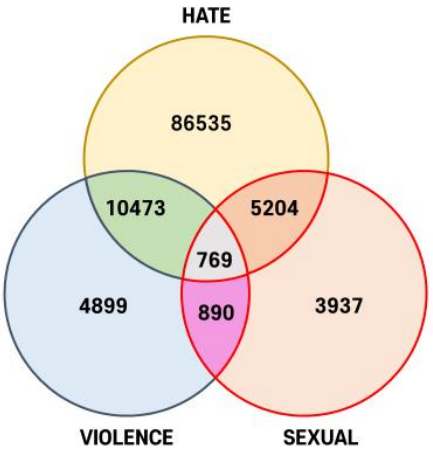
[월 주차 프로젝트 수행 일지]

프로젝트 주제	악성댓글 감정분석		
프로젝트 팀명	Writer's Warden	프로젝트 팀원	김나래, 김성훈, 김은지, 이동근, 장유림

※ 프로젝트 수행 일지는 매 주 금요일 17:00 전까지 최종본을 저장해주세요

1. 프로젝트 수행 계획 및 현황

김나래	1. 데이터 탐색 2. 형태소 분석 3. 혐오사전 구축	< 데이터 정제 >  ● 축소화 과정에 대한 상세 근거 마련을 위해 워드 클라우드 외 다른 방법으로 데이터 탐색 ○ 탐색 I. 분류 카운트 기반 ○ 탐색 ii. 비윤리 강도 점수 분포 ○ 워드 클라우드 ● 결론 :

<table>
<tr>
<td>김은지</td>
<td>1. 데이터 탐색
2. 혐오사전 구축</td>
<td rowspan="2">○ 가장 조합이 많고 악플의 심각도가 평균보다 낮은 '비난', '혐오', ' 차별'을 한 그룹으로 묶고, 텍스트 분석에 따라 '선정' 그룹, '욕설' + '폭력' + '범죄' 그룹으로 라벨을 축소 하고자 했다.
○ 악플의 분류를 3개로 축소하고자 함
(분류의 이름은 팀원들끼리 정한 정의에 따라 부여).
■ 혐오 : 비난(CENSURE), 혐오(HATE), 차별(DISCRIMINATION)
■ 선정 : 선정(SEXUAL)
■ 폭력 : 욕설(CRIME), 폭력(VIOLENCE), 범죄(CRIME)
● 근거1. 카운트 기반 결과
○ '비난'을 제외하고 모두 '단독'보다는 '중복'으로 함께 사용되었다.
○ (중복 2개)의 경우 '비난'과의 조합이 가장 많았다.
○ '혐오' + '차별'(1961) > '폭력' + '범죄'(497) > '혐오' + '폭력'(477) > '선정' + '범죄'(470) > '혐오' + '선정'(402) 순으로 많았다.
○ (중복 3개) '차별' + '혐오' + '선정'(122) > '차별' + '혐오' + '폭력'(116) > '범죄' + '혐오' + '폭력'(100) 순으로 많았다.
● 근거2. 악플의 심각도 기반 결과
○ 악플 분류의 중복개수가 많을수록 심각도가 크다고 할 수 있다.
○ (중복 포함) '범죄' > '폭력' > '선정' > '욕설' > '혐오' > '차별' > ' 비난' 순으로 악플의 심각도가 높았다.
○ (단독) '선정' > '폭력' > '욕설' > '범죄' > '혐오' > '차별' > '비난' 순으로 악플의 심각도가 높았다.

● '선정'을 따로 분류한 이유는
1. '폭력' + '범죄'(497)의 조합이 더 많았고, 3개의 조합에서도 '폭력' + '범죄' + alpha가 많았다.

2. '폭력' + '욕설'(166)이지만, 3개의 조합에서는 '폭력' + '욕설' + alpha의 조합이 더 많았으며,

3. 공통으로 쓰이는 형태소를 분석했을 때, '선정'은 타 분류에 비해 단독으로 쓰이는 형태소의 수가 많았다.
●

< 형태소 분석 >
● 신조어 문제 해결을 위한 user_dic 구축
- Kiwi와 Okt의 필요한 품사만을 추출하여 첫번째 전처리 과정을 거침
- Kiwi와 Okt의 추출 결과를 비교하여 차집합을 만들어 비교
- 가용 불가능한 단어의 특징을 찾아내어 제거
- 최대한 축소된 단어리스트를 직접 분류해 신조어 리스트 구축
- Okt에서 이전에 만들었던 리스트와 합쳐 Kiwi의 User_dic을 구축
- 이전에 사용했던 문장을 User_dic을 이용한 Kiwi를 통해 품사를</td>
</tr>
<tr>
<td>김성훈</td>
<td>1. 데이터 탐색
2. 형태소 분석</td>
</tr>
</table>

<table>
<tr>
<td>장유림</td>
<td>1. 데이터 탐색
2. 혐오사전 구축</td>
<td rowspan="2">추출하여 신조어 문제 해결

< 혐오사전 >
● GloVe 워드 임베딩 이용
- 선정, 폭력, 혐오 별 명사 리스트를 글로브 모델에 학습시켜 유사도 기반 단어 추출
- 선정, 폭력, 혐오 별 혐오 사전 구축
● Fast Text 함수 이용
- 선정, 폭력, 혐오 분류별로 쓰인 명사 리스트 확보
- 선정, 폭력, 혐오 단어사전의 키워드 추출
- 선정, 폭력 혐오 세가지의 혐오사전 구축
● 혐오사전 구축 과정
- 각 분류에서만 쓰였던 형태소를 추출하여 유사도 높은 키워드들 검색
- 각 검색된 키워드들이 포함된 문장을 찾고 그 문장의 분류를 카운트
- 카운트 정보를 기반으로 하여 각 키워드들을 세 개의 혐오사전에 할당

< 모형 - KOBERT >
● 이슈 : accuracy의 진행속도가 매우 더딤
-> batch size를 늘림
● 분석 1. epoch 6번 할당 시 56% / 용량문제로 중단
● 분석 2. (현재 진행중) epoch 4번 89%
● 모델 저장 후 Unsmile에 대한 학습을 재 진행할 예정

<웹 어플리케이션>
<Node.js>
::::: API 관련 개선사항
● Youtube Live Streaming API를 활용해 livechatID, 메시지 내용 뿐만 아니라, 입력시간을 가져올 수 있게 되었습니다.
::::: Python 모듈 관련 개선사항
● 메시지 정보를 비동기 방식으로 실시간으로 Python 모듈에 보낼 수 있게 되었습니다.
● Python 모듈의 표준 출력을 받아와 log에 출력할 수 있게 되었습니다.
::::: 기타 기능 관련 개선사항
● Unsmile data에서 뽑은 랜덤 문장을 Youtube Live Chat으로 보낼 수 있게 되었습니다.

<Python Module>
::::: 실행환경 관련 개선사항
● 서버 컴퓨터의 환경을 jupyter notebook과 동일한 Python 3.9로 변경했습니다.
● jupyter notebook의 가상환경과 동일한 패키지를 직접 설치해 다른 경로에 저장된 Python 모듈이 이를 import 할 수 있게 되었습니다.
::::: Node.js 통신 관련 개선사항
● Node.js에서 호출을 통해 Python 모듈 스크립트가 실행되며, 직접 종료하기 전까지 계속 모듈이 구동되도록 파일의 작동구조가 변경되었습니다.
::::: 내부 프로세스 관련 개선사항
● Google Colab에서 학습시켰던 LSTM 모델을 불러와 Youtube Live Chat에 올라온</td>
</tr>
<tr>
<td>이동근</td>
<td>웹 어플리케이션 작업 진행

Python 모듈 작업 진행</td>
</tr>
</table>

		채팅 데이터를 실시간으로 분석해 표준출력으로 내보내는 기능이 추가되었습니다 . **<주말, 연휴, 차주간 예정사항>** (김나래) 팀원의 KoBERT 모델 전이학습이 끝나는대로 Pytorch를 import 해와 LSTM 대신 KoBERT 모델로 텍스트 분류를 진행할 예정입니다. Unicode, UTF-8, EUC-KR 등 다양한 인코딩 종류가 있는데, Python 모듈에서 자꾸 한글 깨짐 현상이 발생합니다. 이부분과 관련해 조치가 필요합니다.

2. 강사님 피드백	
XX 반 **XXX 강사님**	