

ABUSE



빅데이터를 활용한

마약 및 불법 의약품 거래 게시물 판별

목 차

주제선정
및 필요성

데이터처리

분석모델

흐름도

응용가능성
및 확장성

Ice @Ice08278956
 #아이스 #작대기
 #시원한술 #정품
 #아이스 #팔아요
 #떨 #팔아요 #필로폰
 무조건 안전 문의만주세요!
 칼담해드려요
 몇몇 적은 인원으로 평생 동반자
 되실 확실하신 사장님들구해요 !!

텔레 ice12

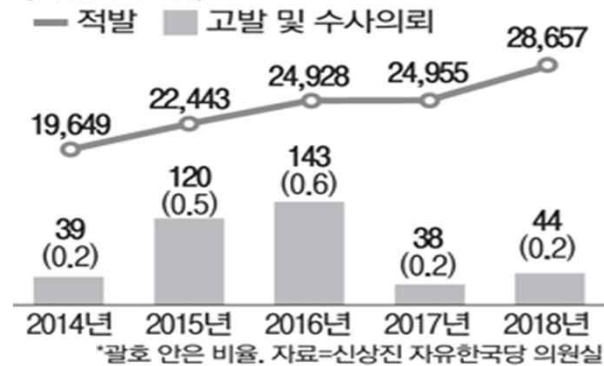
돌아온0.3 @2R4Ee4pWEaWKe7j 3d
 텔 icevip50 텔 icevip50
 #아이스
 #작대기
 #빙두
 #크리스탈
 #차가운술
 #시원한술
 형님~누님들~!!
 돌아온0.3입니다
 vip50명 모시고 장사 하겠습니다
 욕심 버리고 50분 채워지면 광고 중
 단 합니다
 선드랍은 문의 주세요~
 서울 경기 가능 합니다
 텔 icevip50 텔 icevip50



[무작위 대상 높은 접근성]

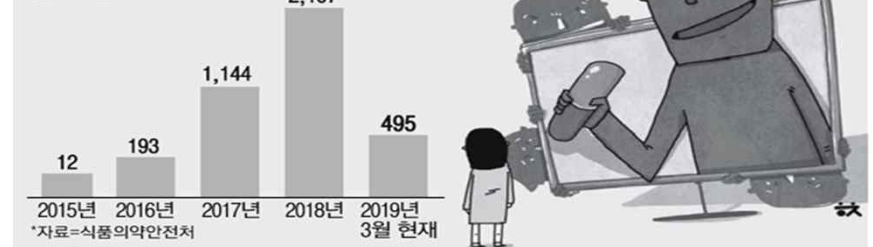
온라인 의약품 불법유통 적발 및 고발

(단위=건 · %)



낙태유도제 온라인 불법판매 적발건수

(단위=건)



[점점 늘어나는 불법 의약품 거래]

- 모델의 학습
- SNS 크롤링
 - a) 정상 게시물_마약옥수수
 - b) 불법 게시물_정품짜대기

활용 도구



Python

Pandas
Request
BeautifulSoup4
Selenium



데이터베이스

CSV

데이터 수집 단계

The diagram illustrates the process of finding a tweet through a search. It shows a search result for the name '이소은' (Lee So-eun), which leads to a specific tweet. The tweet's content is a list of hashtags and a link to a Twitter search page. The source code of the tweet is shown on the right, with a red box highlighting the 'hashtag' and 'link-complex' attributes.

ID	HashTag	Text
thdmskr	작대기,작대기정품 구입,...	최고의품질,신용거 래,...
bigice88	아이스,얼음,빙 드,...	신용과 품질로...
Dnadept_bot		아이스-빔!!
wjajtajt		마약김밥 처음먹어 ...

SNS로 키워드 소스를 제공받음에 있어
나타나는 형태가 많아 전처리가 어렵다

프로포폴

우유주사
건망증 우유
포폴
○ ㄱ 토미데이트
저지방우유

정품

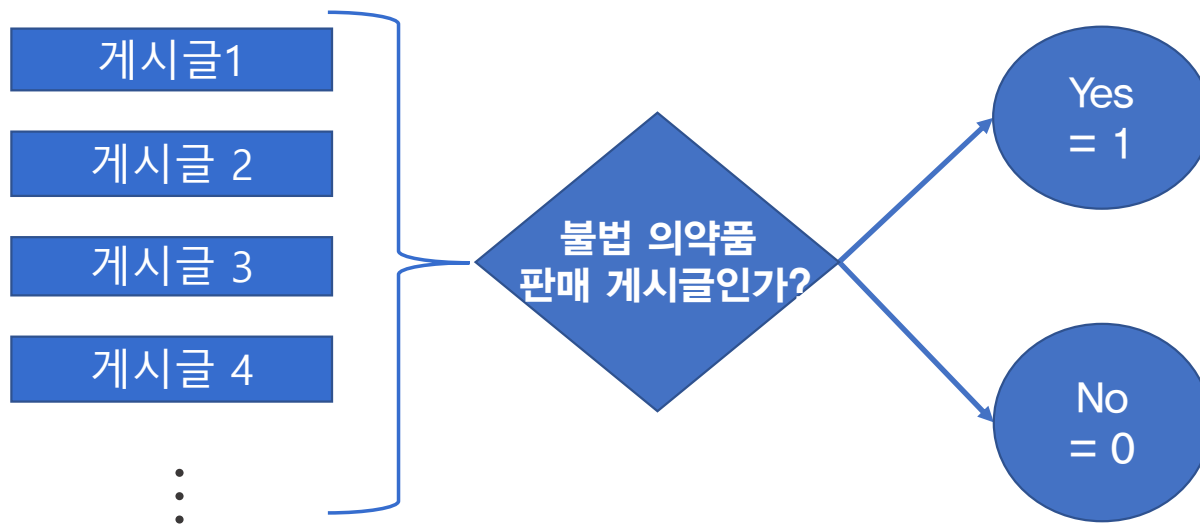
정품
정품입니다
정품
정품팜

낙태약

미프진
미프진
Mifeprex
유산유도제
낙태약

동의어 사전(Syn Source) 작성 을 통한 데이터 관리

게시글 라벨링을 통해 지도학습 훈련데이터 생성



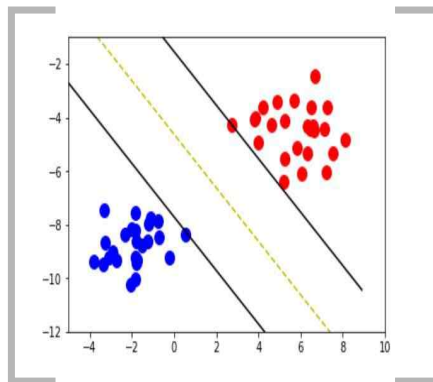
〈훈련데이터를 위한 Labeling 작업〉

ID	HashTag	Text	Is_illegal
thdmskr	작대기,작대기정품구입,...	최고의품질,신용거래,...	1
bigice88	아이스,얼음,빙드,...	신용과 품질로...	1
Dnadept_bot		아이스-빔!!	0
wjajtajt	마약김밥, 시장	마약김밥 처음 먹었...	0

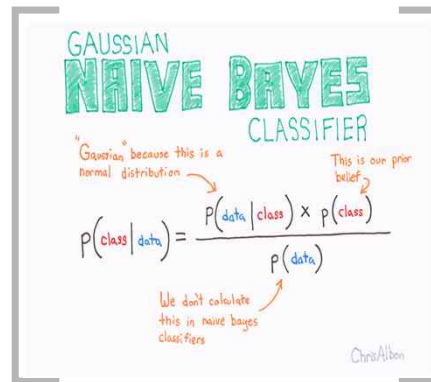
〈Labeling 결과 예시〉

TF-IDF를 응용한 데이터 처리 **모델 선정**

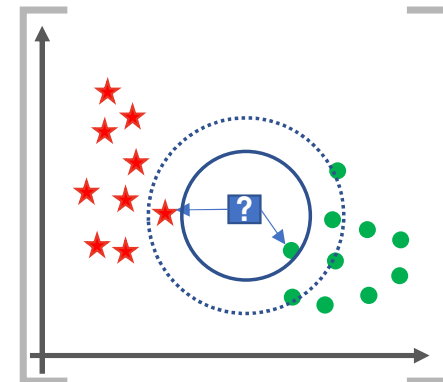
[Support Vector Machine]



[Naïve Bayes]



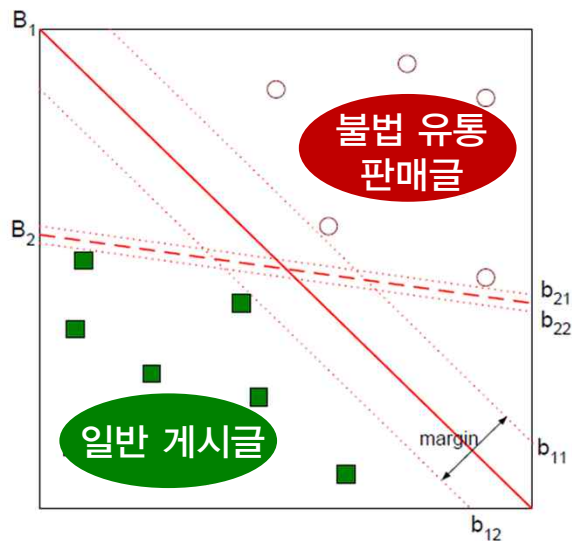
[K-Nearest Neighbor]



양상블 사용 – 모델기반 학습 << 보완 >> 인스턴스 학습

모델 선정에 대한 이유 알고리즘 및 개념

[Support Vector Machine]

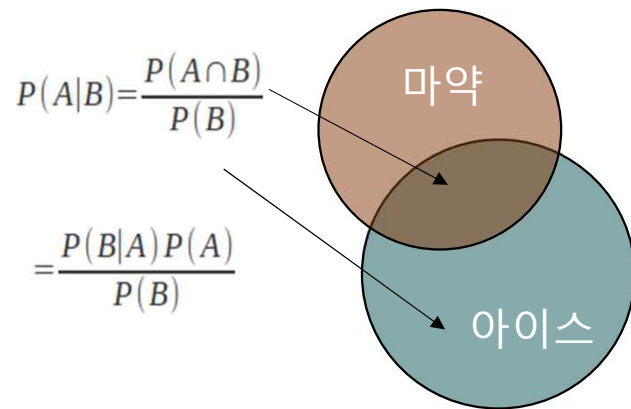


《기계 학습 - 이진 분류》

- 이진 분류(참/거짓 판별)에 뛰어남
- 뛰어난 데이터 처리 속도와 높은 성능
- 고차원 데이터 처리에 높은 정확성을 보임

모델 선정에 대한 이유 알고리즘 및 개념

[Naïve Bayes]

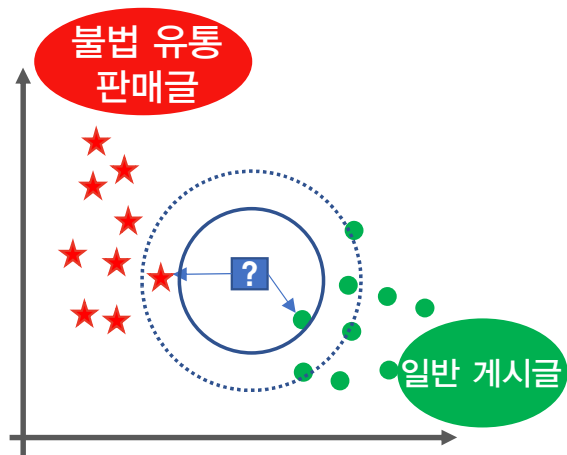


《스팸 필터링 - 조건부 확률》

- 확률 기반의 알고리즘
- 데이터가 증가할수록 높은 정교함
- 고차원 데이터 처리에 높은 정확성을 보임

모델 선정에 대한 이유 알고리즘 및 개념

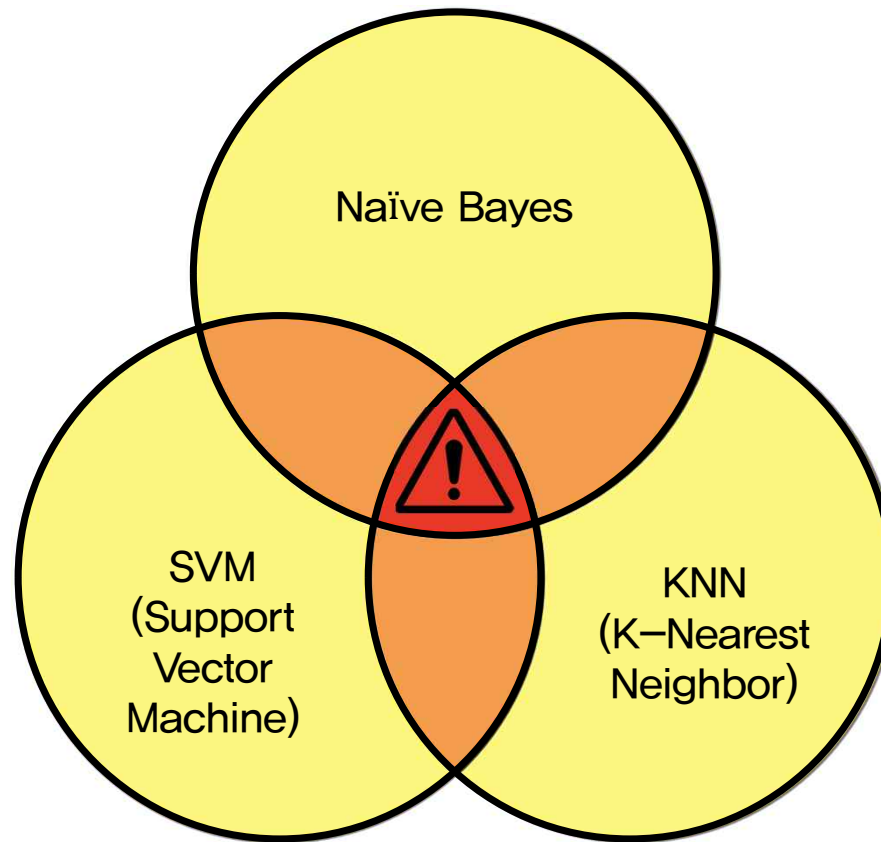
[K-Nearest Neighbor]




《문서분류 - 거리 계산》

- 단어 기반의 데이터 분류
- 인스턴스 기반 학습 모델
- TF-IDF와 함께 효과적인 분류 가능

앙상블 기법
하드 보팅(Hard Voting)



[COUNT]

 (N=3) (N=2) (N=1)

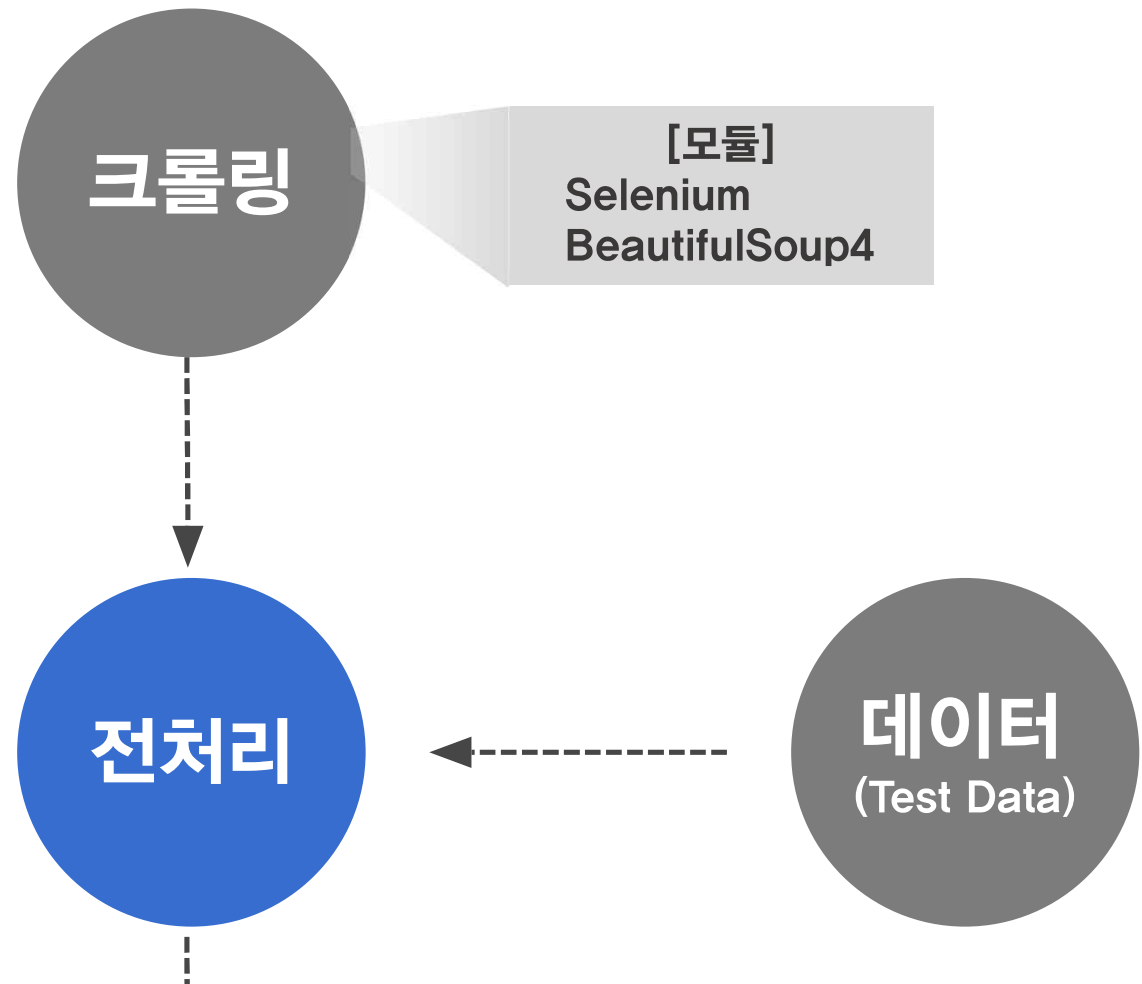
[실시간 모니터링 시스템]

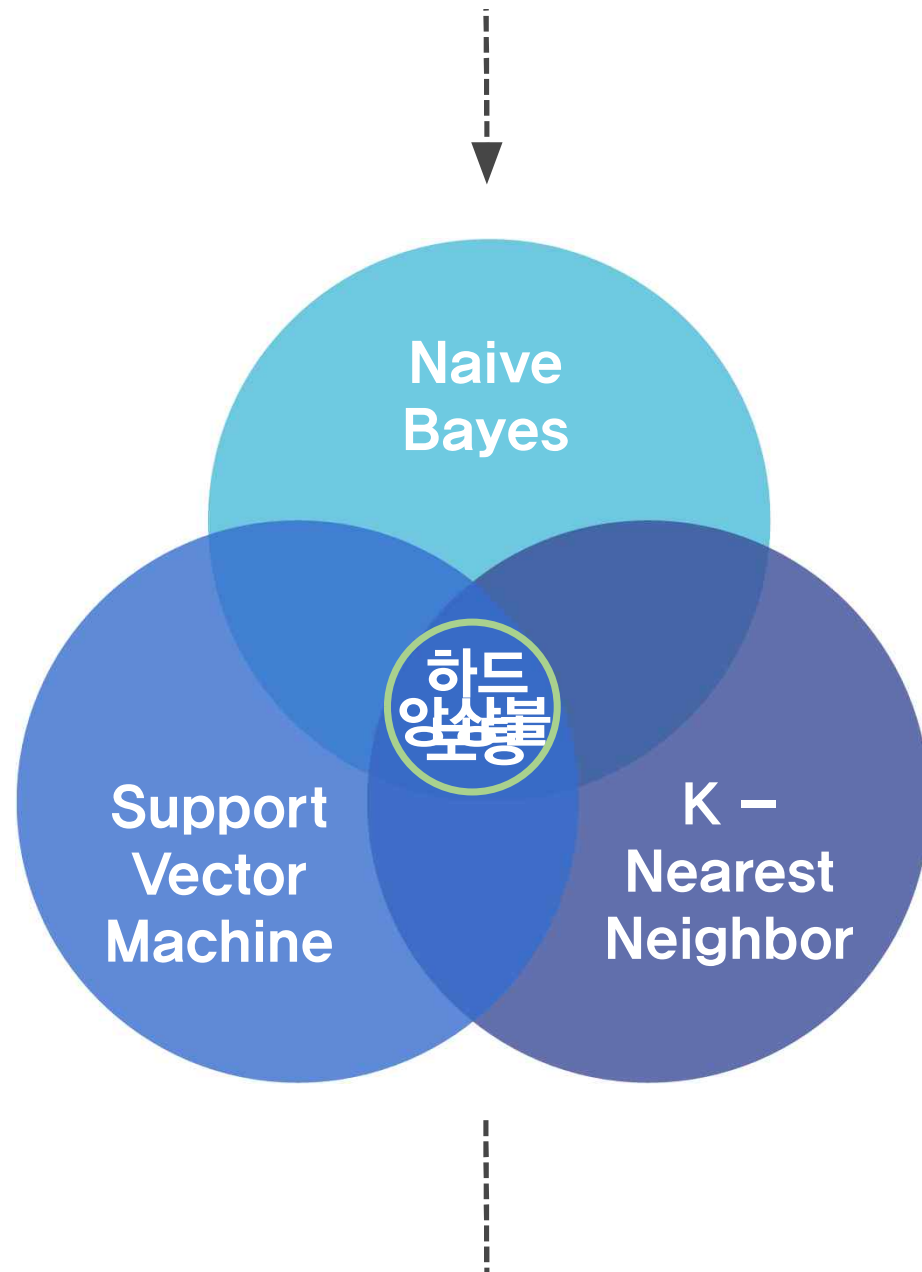


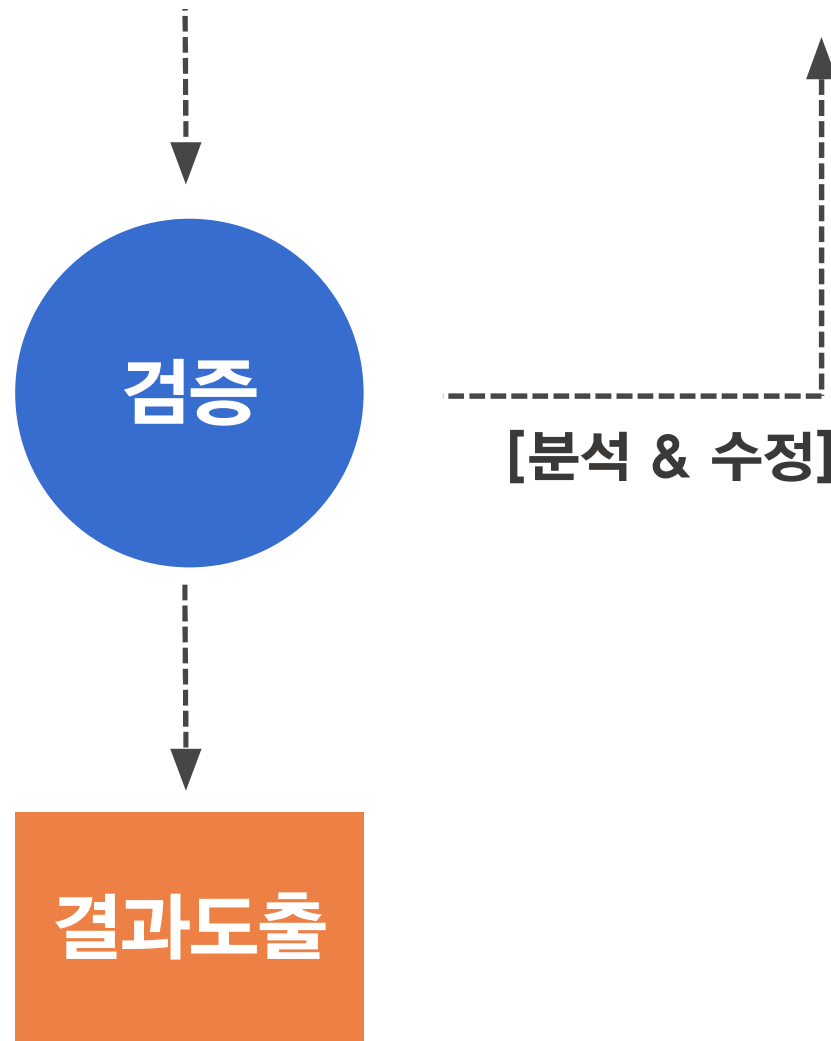
**3개 알고리즘 모두
불법 판매 게시물로 판정**

3개 중 2개 알고리즘
불법 판매 게시물로 판정

3개 중 1개 알고리즘
불법 판매 게시물로 판정







사회기여, 활용성 부분의 서비스 확장성 증대



▶ 비 용 절 감

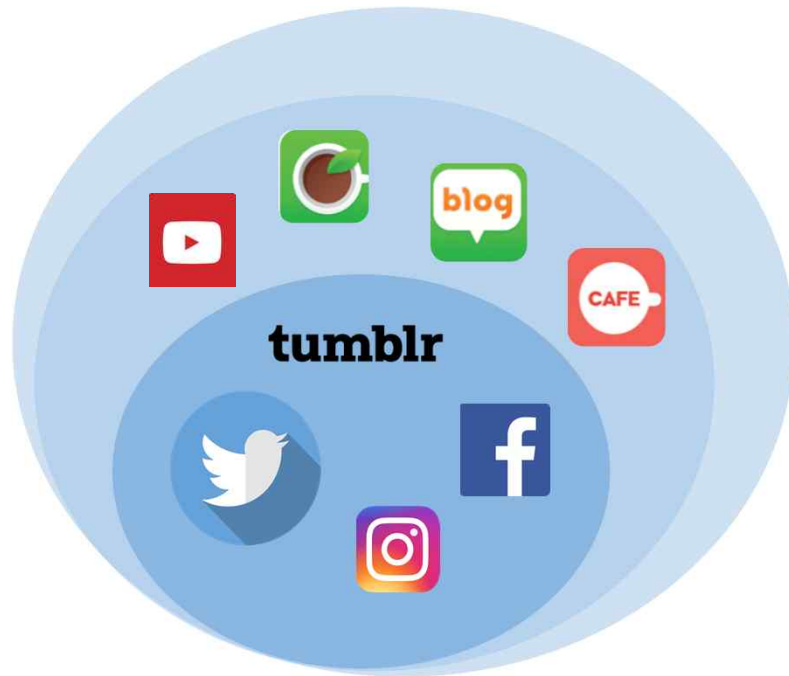
- 모니터링 효율성 증대

▶ 사 전 예 방

- SNS 자동신고
- 글 자동삭제

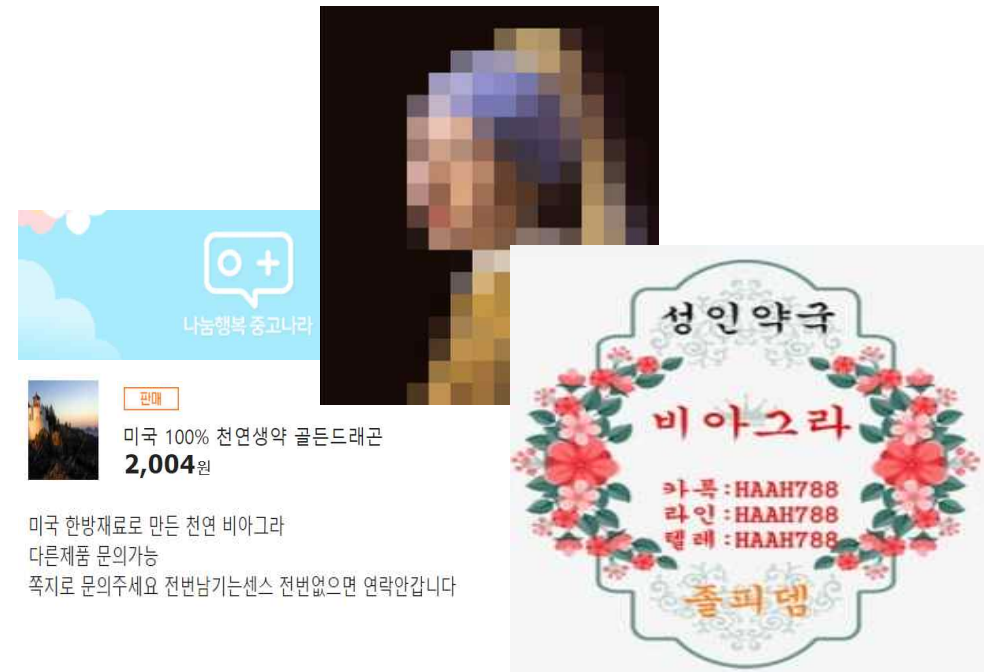
▶ 확 장 가 능 성

- 이미지파일 분석
- 다양한 플랫폼 확장



[온라인 전체로의 감시영역 확장]

SNS 뿐 만 아니라 유튜브, 네이버 카페 (중고나X) 등 온라인 전체로 감시영역 확장 가능



[딥러닝 기반 이미지 인식 적용]

이미지 딥러닝 추가 활용하여 게시물 첨부 이미지 인식

- 수위가 높은 유해사진, 의약품 사진일 경우 불법일 가능성 ↑
- OCR 문자인식 활용 텍스트 추출

THANK

YOU