Automatic Identification of Harmful, Aggressive, Abusive, and Offensive Language on the Web

: A Survey of Technical Biases Informed by Psychology Literature

이상윤 / 박채원

Introduction

- Online Confictual Language (OCL) : overarching category of online language that subsumes all these types also to avoid ambiguity and confusion.
- language = instead of “speech” because the latter implies the spoken nature of the sentence.

- We believe that research on one language might benefit research for another language, and that a precise and organized terminology is needed to improve the quality and applicability of automatic solutions
- This article primarily focuses on psychology, as it provides clear definitions and a diverse set of actionable information.
- discuss the creation of datasets for online conflictual language detection.
- also refrain from focusing on the political aspects of online conflictual languages

- Comparison to previous works
  - their analysis often refers to only few types of languages and only partially explores related disciplines.
  - Our survey substantially departs from previous works precisely by engaging in an analysis of the problem of OCL informed by psychology research

Introduction

Contribution

(1) A set of definitions and properties, and a taxonomy to reconcile the OCL terminology (Section 3). This reconciliation speaks to an increasingly advocated need for conceptual clarity

(2) A discussion of the psychological aspects related to OCLs (Section 4) that uncovers conceptual mismatches with automatic detection works and a reflection on the experimental practices that could contribute to computer science research.

(3) A comprehensive review of the typical data engineering pipelines used for building datasets (Section 5) and of their technical biases (e.g., usage of disagreement metrics for evaluating the annotation quality of subjective OCL) that can be harmful and participate to the low generalization abilities of the systems.

(4) A quantitative review of conflictual language detection models (Section 6) and an analysis of their limitations in terms of performance, leading to the identi#cation of additional biases. Guided by our OCL taxonomy, our work o''ers a principled characterization of di''erences, similarities, limitations, and opportunities in computer science approaches. The lack of features relevant to individual OCL and the integration of social biases are pressing issues, for which future research could draw inspiration from psychology literature and machine learning fairness and explainability literature.

(5) An extensive discussion of open, technical and structural, research challenges, with clear and actionable suggestions for future work inspired by various psychology and computer science domains and informed by our systematic literature analysis (Section 7).

METHODOLOGY AND PAPER COLLECTION

- introduce the methodology employed to achieve the aforementioned contributions, and we explain the procedures followed to collect the computer science and social science papers

- methodology
  - OCL 관련 용어 수집 -> 문헌 탐색&연구 -> taxonomy 생성 -> 연구 challenge 분석

- Paper Collection
  - Retrieval of the List of Terms
    - most comprehensive 한 hate speech survey로부터 시작해 반복적으로 관련 용어 수집(참조된 문헌으로부터)
    - **hate, hateful, toxic, aggressive, abusive, offensive and harmful speeches, profanity, cyberbullying, cyberaggression, flaming, harassment, denigration, impersonation, outing, trickery, exclusion, cyberstalking, flooding, trolling, discrimination.**
  - Retrieval of Psychology Papers
    - 다양한 language를 연구하거나 language 인식에 영향을 미치는 변수에 대한 연구만 모음
    - 관련 없는 문헌을 피하기 위해 사회 심리학 field에만 집중. 추가 문헌을 위해 snow approach 사용
    - (OCL keyword)AND (((variable)OR (perception)OR(de#nition)OR(judgement)) AND((web)OR(online)OR(internet)) AND (source:“social Psychology”) -> google scholar에 이렇게 검색함.
  - Retrieval of Computer Science Paper
    - list of term 과 몇개의 키워드(ex/ filtering, crowd, crowdsourcing 등)를 and 해서 query
    - 2018년 말에 수집되었고, 2019년, 2020년 work가 추가됨. CS field에 제한해서 검색
    - 겹치는 것, trolling, spamming에 관한 것 필터링. online game에서의 유해한 행동에 대한 문서들도 필터링.
    - 모은 논문에서 참조된 논문들을 모으고 다시 필터링 함.

TERMINOLOGICAL MISMATCH: ENTANGLED DEFINITIONS

- 사회과학, 특히 심리학에서 OCL language가 어떻게 정의 되었는지 분석하고, 정의를 조정하고 분류체계를 만든다.

  - Definitions of OCL : retrieved from a psychology dictionary and psychology literature
    - **“Hate”** : hate crime과 공통 속성을 갖는다.
      - +) cyberhate - namely, online messages demeaning people on the basis of their race/ethnicity , gender, national origin, or sexual preference
      - 이 정의는 명확하게 language의 대상의 타입으로 정의 된다.
    - **“Aggression”** : 논의가 이루어지고 있다.
      - aggression을 “behavior that result in physical or psychological harm”으로 정의한다면우리는 피해를 입히려는 의도가 있었는지 없었는지에 대한 question을 던져야한다.
    - **“Bullying”** : “physical, verbal, or psychological intimidation that is intended to cause fear, distress, or harm to the victim” [219] with “the repetition of the behaviour over a period of time and the relational asymmetry between bully and victim”
    - **“Discrimination”** : “harmful actions toward members of historically subordinated groups because of their membership in a particular group. [...]Discriminatory behaviors are carried out based on personal prejudices or stereotypes about members of a specific group.
    - **“verbal sexual harassment”** : “judgments of appearance, obscene and euphemistic statements about sexual receptivity, and remarks belittling the competency of one’s gender

TERMINOLOGICAL MISMATCH: ENTANGLED DEFINITIONS

- Reconciled Definitions
  - CS에서 OCL관련 용어가 항상 정의돼있지 않거나 모호하다.
  - 용어에 대한 제대로 된 정의를 하는 연구가 몇 없고 그조차도 완벽하지 않다.
  - dict에서 모은 정의는 정확하지도 않고 부합하지도 않아서 정의를 조정하고 분류법 또한 정의하고자 함.
  - Methodology
    - 컴퓨터 과학과 사회 과학 모두 정의가 존재할 땐 사회과학의 정의 사용. 더 깊게 연구되었기 때문에
    - 컴퓨터 과학에만 정의가 존재할 땐 논문에서 나타나는 빈도에 따라 single definition을 select 한다.
    - 용어에 대한 직관과 가장 비슷한 정의 선택.

Table 7. Selected Definitions of *OCLs*

Language	Definition
Offensive	Communication which attacks persons on some of their characteristics, most often with rude language. (combination of References [58, 209, 220, 285, 296])
Hateful speech	Speech which contains an expression of hatred on the part of the speaker/author, against a person or people, based on their group identity. [233]
Hate speech	Language used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group. (from Reference [145], similar to References [7, 79, 102, 173, 174, 209, 248, 282, 304])
Aggression	Intention to harm. [81, 148, 209]
Cyberaggression	Online aggressive behavior with intention to harm. [53, 102, 130, 216]
Cyberbullying	Willful and repeated harm inflicted to an individual through the medium of electronic text. [6, 71, 72, 77, 85, 178, 179, 185, 233, 256, 258]
Flaming	Online fights using electronic messages with angry and vulgar language. [247]
Harassment	Repeatedly sending nasty, mean, and insulting messages to intentionally annoy others. [298]
Denigration	Dissing someone online. Sending or posting gossip or rumors about a person to damage his or her reputation or friendships. [247]
Impersonation	Pretending to be someone else and sending or posting material to get that person in trouble or danger or to damage that person's reputation or friendships. [247]
Outing	Sharing someone's secrets or embarrassing information or images online. [247]
Trickery	Talking someone into revealing secrets or embarrassing information or images online. [247]
Exclusion	Intentionally and cruelly excluding someone from an online group. [247]
Cyberstalking	Repeated, intense harassment and denigration that includes threats or creates significant fear. [247]
Flooding	Repeatedly entering the same comment, nonsense comments, or holding down the enter key for the purpose of not allowing the victim to contribute to the conversation. [29]
Trolling (baiting)	Intentionally posting comments that disagree with other posts in the thread for the purpose of provoking a fight, even if the comments don't necessarily reflect the poster's actual opinion. [29]
Abusive	Any strongly impolite, rude or hurtful language using profanity, that can show a debasement of someone or something, or show intense emotion. (Reference [103], close to References [1, 133, 155])
Toxic	Rude, disrespectful, aggressive comment likely to make somebody leave a discussion. [292]
Hate	Expression of hostility without any stated explanation for it. [101]
Discrimination	Process through which a difference is identified and then used as the basis of unfair treatment. [101]
Profanity	Offensive or obscene word or phrase. [101]
Harmful	Text which has a negative effect on somebody. (proposed based on dictionaries)

TERMINOLOGICAL MISMATCH: ENTANGLED DEFINITIONS

- Reconciled Taxonomy
  - 기존의 분류법은 차이점을 강조하지만 관계는 명확히 만들지 않음.
  - concept의 카테고리화를 위해 공통 속성을 제공하고 분류법을 이끌어냄.
- Methodology
  - 7개의 속성 - 독립적인 카테고리화를 위해서.
  - 해당 속성을 concept가 필수적으로 포함할때 yes 표시, 필수적이지 않으면 no

properties : Intention, Behavior, Specific focus, Emotion of the author, Language, Target, Effect

main group : Aggresion. Offensive, Abusive, Harmful

본질적으론 겹쳐있기 때문에 완전히 독립적이진 않다.

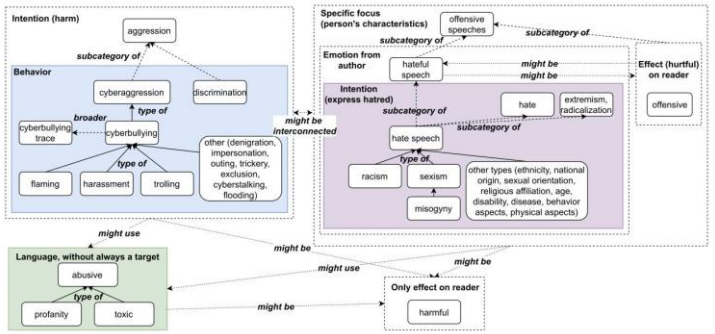


Table 8. Analysis of the *OCL*

Concept	Intention *Hatred*	Intention *Harm*	Behavior	Specific focus *Other*	Specific focus *Character.*	Emotion (hatred)	Language	Target	Effect
offensive	N	N	N	N	Y	N	N	Y ((type) person)	Y
hateful speech	N	N	N	N	Y (stereo)	Y	N	Y (person, group)	N
hate speech	Y	N	N	N	Y (stereo)	Y	N	Y (person, group)	N
aggression	N	Y	N	N	N	N	N	Y	N
cyberaggression	N	Y	Y	N	N	N	Y	N	
cyberbullying	N	Y	Y (repetitive aggression)	N	N	N	N (often aggressive)	Y (power imbalance)	N
flaming	N	Y	Y (fight)	N	N	N	Y (abusive)	Y	N
harassment	N	Y	Y (repetition)	Y (offense target)	N	N	Y (abusive)	Y	Y (annoy)
denigration	N	Y	Y (damage reputation)	Y (gossip, rumor)	N	N	N	Y	N
impersonation	N	Y	Y (pretend to be the target)	N	N	N	N	Y	N
outing	N	Y	Y (sharing target's secret)	Y (secret)	N	N	N	Y	N
trickery	N	Y	Y (forced info. sharing)	Y (private info.)	N	N	N	Y	N
exclusion	N	Y	Y	N	N	N	N	Y	Y (exclusion)
cyberstalking	N	Y	Y (repeated)	Y (offense target)	N	N	N	Y	Y (threat, fear)
flooding	N	Y	Y (repetitive posting behavior)	N	N	N	N	Y	Y (exclusion)
trolling	N	Y	Y (disagreeing posts)	Y (disagree with posts)	N	N	N	Y	N
abusive	N	N	N	N	N	N	Y	N	N
toxic	N	N	N	N	N	N	Y (disre-spectful)	N	Y (leave the discussion)
hate	Y	N	N	N	Y	Y	N	Y	N
discrimination	N	Y	Y	N	Y (difference)	N	N	Y	Y (unfair treatment)
profanity	N	N	N	N	N	N	Y (rude)	N	Y (offense)
harmful	N	N	N	N	N	N	N	N	Y

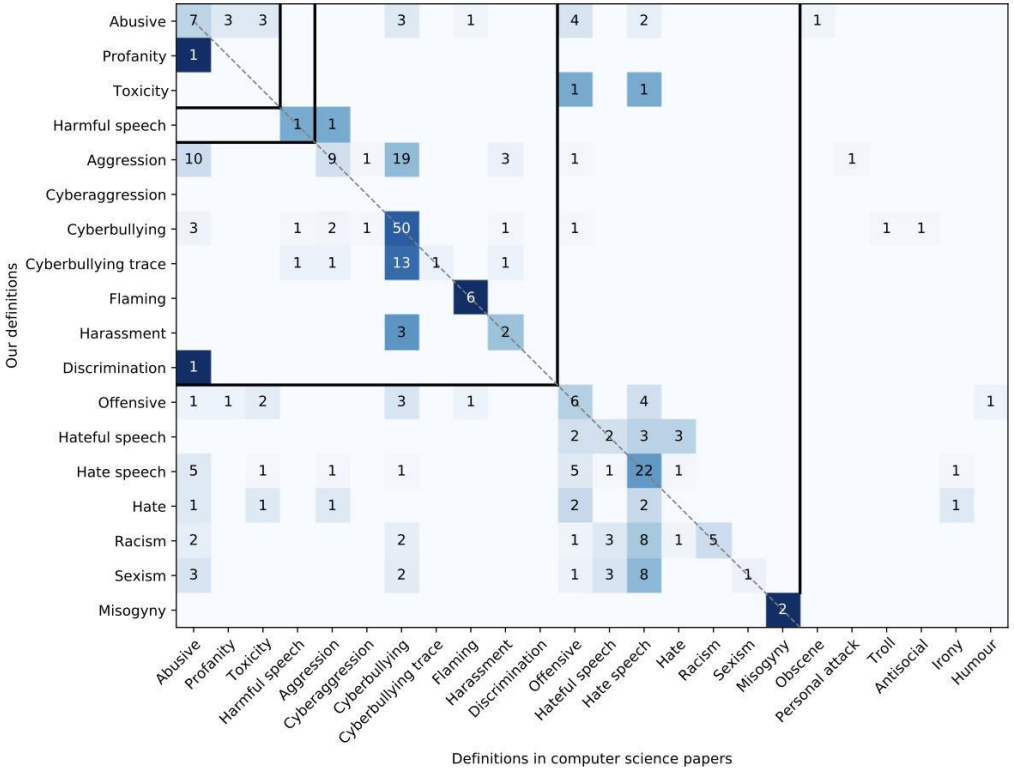
The different *OCLs* are classified along the dimensions of studies identified earlier. They are later clustered (colors) in groups that share similar characteristics, to organize them into a taxonomy. "Stereo" refers to stereotype. ("Y" stands for "yes" and "N" for "no.")

TERMINOLOGICAL MISMATCH: ENTANGLED DEFINITIONS

- Mapping of the Computer Science Literature into the Revised Taxonomy
  - 용어가 원래 CS에서 쓰였던 방법과 새로운 분류법에 따른 정의를 mapping
  - CS에서 용어에 대한 정의 합의가 되지 않았다. -> 잘못된 식별을 낳을 수 있다.
  - 더 넓은 개념을 연구하는 데이터셋의 부족때문일 수 있다.

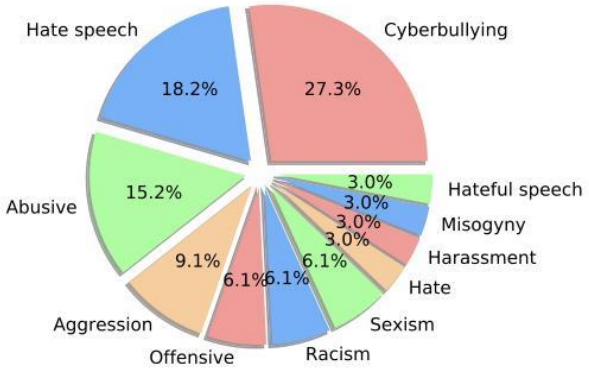
이 논문에서 정의된 정의를 기존 paper에 적용해보면 많은 concept들이 혼동되어 사용되고 있었다.

예시)
In 10 papers, the word *abusive* is used to refer to *aggression*,
while *offensive* is confused, respectively, with *sexism, racism, cyberbullying, aggression* (1 time), *hate, hateful speech* (2 times), *hate speech* (5 times), *abusive language*(4 times)

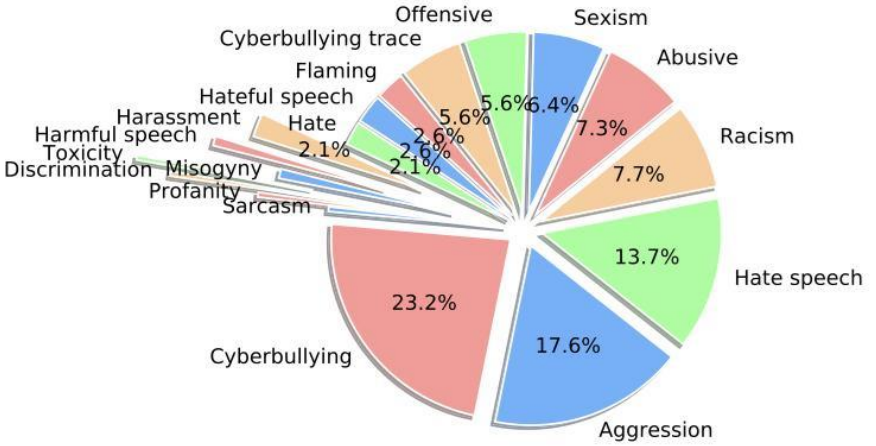


TERMINOLOGICAL MISMATCH: ENTANGLED DEFINITIONS

- Distribution of Works on OCL concepts



(a) *OCL* concepts in studies.



(b) *OCL* concepts in publications that develop solutions for classification tasks.

미디어에서의 language의 인기와, finer-grain에 대한 이해 부족으로 인해 불균형이 발생한다.
finer-grain concept을 감지하는 것은 더 정확하고 모듈식의 필터링을 가능하게 한다.

CONCEPTUAL MISMATCHES TOWARDS TECHNICAL BIASES

- Semantic knowledge from Psychology / 심리학 문헌에 확인된 OCL 인지에 영향을 주는 요소들
  - Internal characteristic of the observer
    - ex/ **성별**에 따라 느끼는 aggressiveness율이 다르다
    - ex/ **성별**과 **자유주의 성향**이 hate speech가 얼마나 유해하게 들리는지의 정도에 영향을 준다
    - 이것은 올바른 변수의 중요도를 설명하기 위해 명확한 정의의 중요성에 대해 강조한다.
  - Sentence content and context
    - profanity를 community마다 다른 정도로 기분나쁨을 인지함. -> context
    - 그룹보다 개인에 대해 비난하는 것을 더 offensive하게 생각한다. -> content
    - 즉 어떤 language의 context와 content에 따라 offensive의 정도를 다르게 느낀다.
  - 세 타입의 변수는 암시적으로 language의 finer-grained 특성을 포함한다.
  - 이러한 요소 때문에 정의를 제대로 하는 것이 매우 중요.

Category	Variable	Measure	Paper
Observer	Gender	Question	[66, 67, 93, 120]
Observer	Ethnicity	Question	[66, 67, 289]
Observer	Education	Question	[66, 67]
Observer	Age	Question	[66, 67]
Observer	Liberalism inclination	Question (scale)	[93]
Observer	"Individuals' attributions of intent", angry and anxious dispositions	Not investigated	[120]
Observer	Sense of mastery, self-esteem	Question	[204]
Observer	Frequency to which people are subject to racial prejudice, "beliefs about the appropriateness of expressing racial prejudice"	Question (scale)	[186, 289]
Observer	Membership esteem to the offended group	Question (scales)	[37]
Context/Content	Targeted group or person	Scenario	[37, 66, 67, 126]
Content	Category of hate speech	Info in dataset	[126]
Content	Prejudice, sentence properties	In the dataset	[66, 86]
Context	Public or private sentence	Scenario	[66, 67]
Context	Received response to the language	Scenario	[66–68]
Context	Author, its characteristics, race, gender	Scenario	[70, 207]
Context	Hierarchical level of perpetrator and victim	Question	[265]
Context	Internet community	Info in dataset	[253]
Context	Social status of a group	Question	[126]

CONCEPTUAL MISMATCHES TOWARDS TECHNICAL BIASES

- Context of application of the systems.
  - platform마다 사용되는 OCL이 다를 수 있다.
  - 이러한 특성이 누군가의 OCL 인지에 영향을 끼칠 수 있다.
  - 이러한 효과를 이해하면 시스템을 사용할 수 있는 컨텍스트의 범위를 지정할 수 있다.
  - 정부나 소셜 미티어의 법률 및 규정은 감지될 OCL의 타입을 더 제약한다. language의 특성에 더 집중함.

- Hard technical requirements for the application
  - 기술적 어려움이 존재. ex/ OCL 포스트가 제거되어야 하는 시간 제한이 요구된다.
  - 이는 OCL의 특성에 필수적으로 영향을 주지 않는 반면 탐지 파이프라인에 제약을 부과한다.
  - 시간은 정확도와 tradeoff 관계 (scalability vs accuracy)
  - 그러나 이러한 요구는 문헌들에서 잘 드러나지 않는다. 정확도에 초점을 맞춘다.
  - 4%만이 scalability 언급, 6%만이 full system의 생성에 대해 얘기했다.
  - 지속적으로 데이터 세트를 수집하는 시스템은 극소수.

CONCEPTUAL MISMATCHES TOWARDS TECHNICAL BIASES

- Computer Science Framing of OCL
  - OCL 감지 시스템 설계에서 기술적 편향으로 해석되는 개념적 불일치를 식별
  - 컴퓨터 과학 출판물의 탐지 작업공식
  - 편견에 대한 작업의 개요 제공
  - Framing of Automatic Detection Task
    - 분류 작업이 어떻게 짜여졌는지
    - 작업에서 사용되는 클래스와 entities의 다양성을 보여준다.
      - entities - 치우쳐져있다.

	Aggression	Offensive	Abusive	Harmful language
Media sessions	6	0	0	0
Sentence	83	75	12	1
User	13	1	1	0
Words	3	0	1	0

classes - class 2,3개가 일반적이고 class가 4~13의 경우는 더 감지하기 어려운 OCL 언어의 강도를 나타낸다.

CONCEPTUAL MISMATCHES TOWARDS TECHNICAL BIASES

- Main Bias Concerns : Bias유형
  - Inherent contextual biases
    - 실제 observers는 어노테이터와 동일하지 않다. 때문에 연구는 실제 유저의 인식에 맞지 않다. 대화 맥락이 조사되지 않음
  - Biases related to the online context of the systems
    - 플랫폼의 맥락적 특성이 언급되지 않는다.
    - 플랫폼간 사용자의 OCL에 대한 다양성 인지는 연구 안됨.
    - OCL을 쓰는 유저만 씀. 이로인해 유저를 식별하는 것으로 인식해서 모델을 과대평가한다.
  - Discrimination-related biases
    - 편향은 모두 최종 사용자의 속성과 애플리케이션의 자연어로의 번역에 의존한다.
    - 사용자의 다양한 하위 집단에 대한 시스템 성능을 비교하여 식별된다.
    - 데이터 세트의 다양한 특성의 불균형(ex/ 남성 작성자가 작성한 문장이 더 많음)

  - 편향을 설명하는 컴퓨터 과학 작업은 아직 모든 종류의 관련 문맥 및 의미 정보를 포함하지 않는다.
  - 이 정보를 고려하지 않음으로써 발생하는 기술적 편향을 식별한다. 다음에서.

DATASET construction for the detection of OCL

- OCL 탐지 시스템에 사용되는 dataset과 pipeline을 분석
  - 실험이 수행된 194개의 publication 중 33%만이 이미 존재하는 dataset을 사용했다.
  - 시간과 비용이 들더라도 dataset을 생성할 필요가 있다.

- Data Sample Collection
  - 웹의 다양한 소스에서 수집됨
  - 트위터는 대중성과 데이터 확보의 용이성으로 많이 사용되지만 다른 소셜 미디어는 덜 사용됨.
  - 뉴스 웹사이트 같은 다양한 사이트는 스포츠 정치와 같은 하나의 주제를 전문으로 함.
  - 영어 데이터가 압도적으로 많음.

Table 3. Dataset Sources Distribution

Data source	Count
Twitter	98
Formspring	18
News site	16
YouTube	14
MySpace	14
Forum	13
Wikipedia	12
Facebook, individual or group conversations	11
Instagram	9
Yahoo	8
Other content-sharing social media	7
AskFM	7
Website (non social media, e.g., Tumblr, Whisper)	6

Table 4. Datasets Language Distribution

Sample language	Count
English	157
Indonesian	6
Japanese	6
Dutch	5
Spanish	4
Portuguese	4
German	4
Arabic	3
Hindi	3
English-Hindi	3
French	2
Korean	2
Greek	2
Italian	2
Bengali	1
Russian	1
Turkish	1

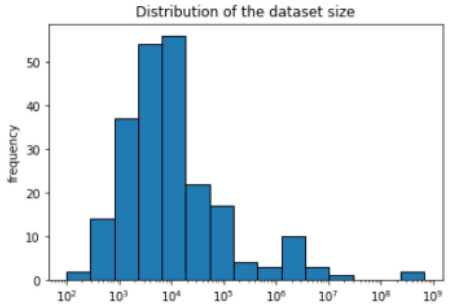
DATASET construction for the detection of OCL

- Data Mining
  - 대부분의 dataset은 욕설, 논란이 되는 정치 사이트, 공격적 언어로 검색된 것이다.
  - 몇 논문은 snowball sampling 또는 해시태그 기반으로 하는 트윗을 먼저 검색한 후 작성자의 모든 트윗을 검색하는 등 변형.
  - OCL을 포함할 가능성이 있는 전체 페이지를 크롤링하여 검색
  - 또는 소셜 미디어 피드를 크롤링, 무작위로 샘플링, OCL 샘플을 최대화하기 위해 키워드 또는 부정적 어휘를 기반으로 한 추가 필터링을 적용함.

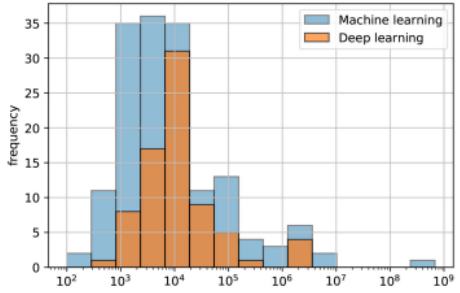
- Introduction of biases
  - 데이터 수집을 위한 매개변수 설정은 dataset을 편향시킨다.
  - 데이터 수집 기간과 같은 매개변수 조차 왜곡을 발생하기도 한다.
  - 심리학에서는 OCL을 분류할 때 문맥을 중요하게 생각하지만 대부분 머신러닝 논문에서는 그러지 못한다.

DATASET construction for the detection of OCL

- Data Processing
  - Data Augmentation (데이터 증대)
    - 기존의 머신러닝 방식보다 딥러닝 방식은 dataset의 요구량이 많다.
    - Oversampling 또는 특정 클래스의 Undersampling, 또는 둘 다 수행하며 SMOTE와 같은 기법을 사용한다.
  - Pre-processing data samples
    - 영어에 대한 표준 형식의 데이터 전처리를 한다.
      - stop words removal : 중단어 제거
      - tokenization : 토큰화
      - stemming : 형태소 분석
      - lemmatization : 표제어 지정
  - Introduction of biases
    - Grondahl 외 의 논문 -> dataset과 동일한 분포를 가진 dataset에서는 잘 수행되는 모델이 다른 dataset에서는 잘 수행되지 않음. but 다른 분포를 가진 dataset을 다시 훈련하면 이는 잘 수행됨.
    - 이 결과는 모델의 구조가 성능의 주요 원인이 아니라 dataset 자체에 모두 자체적 편향이 포함되어 다른 dataset으로의 일반화를 방해함을 보여줌.
- 데이터 증대 및 전처리는 편향을 강화하거나 시작하게 한다.



(a) General distribution.



(b) Distribution per classification method.

DATASET construction for the detection of OCL

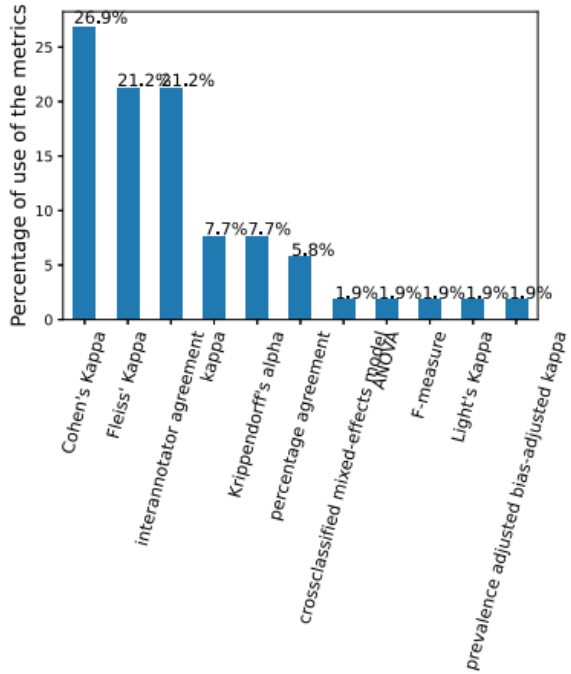
- Data Annotation Collection
  - 기본적으로 여러 annotator의 입력을 집계하여 수집
  - 몇몇 논문은 머신 러닝, data context에서 추론 또는 반지도 학습을 사용하여 레이블을 추론함.

- Set-up of the Annotation Process
  - Annotator에 대한 지침 : 일반적으로 이진분류 + 미정
  - 심리학 문헌에서는 valence scale은 이진분류보다 애매하다고 주장. (감정가 척도)
  - 크라우드 소싱을 사용하는 74개의 논문 중 32% 만이 offensiveness 와 hate speech에 대한 기준을 제시.
    - 명확한 정의를 제공하지 않는 것은 문제이다.

DATASET construction for the detection of OCL

- Data Annotators
  - CrowdFlower 전문가 (23.7%) , 대학생(13.8%) , Amazon Mechanical Turk(15%)

- Annotation aggregation
  - 정보를 이용할 수 있는 50개의 논문 중 49개의 논문이 여러 주석자의 주석을 이진 레이블로 집계함
  - 대부분 다수결 방식을 사용
- Annotation quality control
  - 34.2%의 논문이 고품질의 라벨을 얻기 위해
    - 정확한 정의와 명확한 질문을 사용
    - 주로 3명의 annotator를 사용 (다수결 사용시 홀수가 유리함 + 비용 문제로 4,5명 이상은 지양함)
    - Annotator간의 일치를 측정하여 품질을 평가하기 위한 메트릭으로 Cohen's Kappa, Fleiss's Kappa 등 사용



DATASET construction for the detection of OCL

- Biases in the Annotation Process
  - “The data annotation process introduces various types of biases with each of the design choices”

  - Identification of mismatches
    - 공격적 언어의 dataset을 개발할 때, 공격성에 대한 특정 정의는 문장의 맥락, 작성자의 행동, 판단하는 사람을 보고 문장이 어떻게 인식되는지를 알아야한다.
    - e.g., aggression is “neither descriptive nor neutral. It deals much more with a judgmental attribute”
    - 심리학은 이러한 판단에 문화적 배경이 영향을 미친다고 함.
    - 유사한 예는 대상 그룹의 상태에 대한 인식에 따라 달라지는 그룹 기반 비방의 공격적 인식이다.

- Missing context information
  - 심리학 문헌은 많은 상충되는 language의 경우 sample context가 sample의 인식에 영향을 미친다는 것을 보여준다.
  - 그러므로 대부분 corpus에 context가 포함되지 않는 경우가 많지만 이런 제한을 인정하고 개선하기 위해 노력해야한다.

- Lack of annotator control and information
  - 심리학은 OCL이 주관적임을 강조하며 언어학은 OCL의 해석의 다양성을 보인다.
  - 개인의 특성 (나이, 성별 등)이 라벨링에 영향을 끼친다.
- Simplification of the annotations
  - 주석이 처리되는 방식은 편향을 만든다.
  - 주석을 단일 레이블로 집계하는 것은 주관성을 배제하며 이는 dataset을 특정 유형의 인식, 일반적인 다수의 의견으로 왜곡한다.

Classification Models for the Detection of OCL

- OCL 감지에 사용되는 알고리즘에 대해 설명, 데이터, 선택된 평가 절차에서 추출된 feature에 집중.
  - Implicit biases를 식별하는 것을 목표한다.

  - Features for Classification
    - Types of features extracted from data

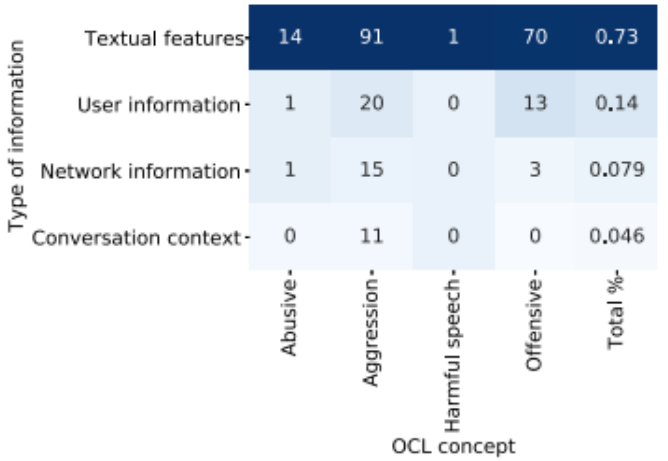


Fig. 9. Type of information used by the classification methods according to the *OCL* concepts.

1. Textual features : 분류 방법에 따라 다르게 표현됨. Word n-gram, Bow(Bag of Words) 및 임베딩은 머신러닝 분류기에 적응된 입력이라 대부분 사용됨.
2. Information about the users : 두번째로 많이 사용하는 정보로, 팔로워 및 친구 수를 기반으로 한 소셜 미디어에서의 사용자의 인기도, 좋아요를 누른 트윗 수에 기반한 사용자 활동, 성별, 나이 및 위치가 포함된다.
3. Information about the network of the users : 사용자가 팔로워들과 얼마나 반응을 하는지를 측정하는 것.
4. Conversation context : 데이터 샘플 주위의 대화, 일련의 질문과 답변, 소셜 미디어에서 text sample과 함께 찾은 이미지 및 해당 캡션 등에 대한 정보

Classification Models for the Detection of OCL

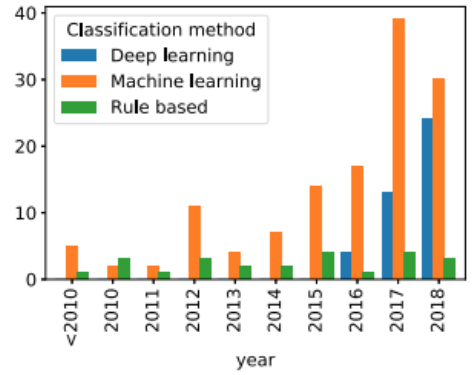
- Feature Selection
  - 분류 성능을 높이기 위해 입력 특징의 차원을 줄이기도 한다.
  - 약 12%가 사용했고 feature 가중치는 SVM점수, Logistic Regression 가중치가 함께 사용되거나
  - 분류기의 출력을 하나의 특징으로 나타내고 이 점수를 기반으로 선택하는 방법을 사용한다.
- Introduction of Biases
  - Feature를 선택하면 모델이 특정 유형의 정보를 사용하도록 자동으로 편향되고 오류가 편중된다.

    - Mismatch with psychology : feature가 가공되는 방식에서, 심리학에선 text를 둘러싼 문맥도 OCL 인식에 영향을 끼침을 보이지만 머신러닝에서는 약 23%의 논문만이 추가 정보를 사용했다.
    - Lack of OCL-dependent features : 여러 실험 연구에서 머신러닝 모델이 서로 다른 OCL 구별을 잘못함을 보였다.
      - hate speech 와 profanity의 구분
      - 각 특정 OCL에 대한 기능의 적응 부족을 보인다.
    - Discriminatory features :단어 임베딩으로 비롯된 특정 feature의 차별적 특성이 있다.
      - Caliskanet 외 는 단어 임베딩의 편향을 측정하기 위해 심리학 테스트를 한 결과, 임베딩이 인간 편향을 재현한다는 것을 보여주었다.
      - Garg 외 는 다른 기간의 corpus로 부터 훈련된 임베딩 텍스트로부터 직업 관련 편향이 포함됨을 보였다.

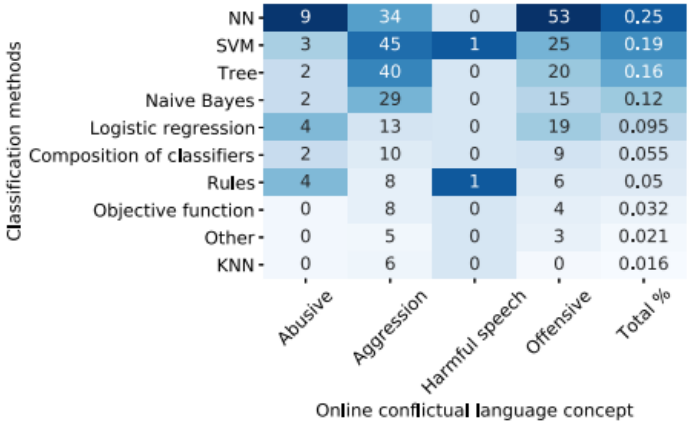
Classification Models for the Detection of OCL

- Methods for Classification
- Overview of the Classifiers
  - 분류 방법의 세가지에 주목한다.
    - rule-base models
    - machine learning models
    - deep learning models

대부분의 머신러닝 논문은 SVM, 트리기반 분류기(Decision tree, Random forest), Naive Bayes Classifier, MLP(다중 퍼셉트론), RNN 및 이들의 조합을 사용한다.



(a) Evolution of the classification methods over years.



(b) Distribution of classification methods per *OCL* coarse-grained concepts. Colours represent the frequencies of methods per concept (white to blue, lowest to highest frequency).

Fig. 11. Quantitative analysis of the classification methods.

Classification Models for the Detection of OCL

- Training Process
  - dataset 수집, 모델 생성의 파이프 라인을 따른다.
  - 지도학습, 반지도학습 수행, feature selection과 분류기 학습을 동시에 진행한다.

- Introduction of Biases
  - Aggregation bias : 집계 편향, 다양한 개별 인구에 대한 단일 머신러닝 모델의 개발 및 적용에 의해 정의된다.
  - 다수결로 결정하는 경우
  - Mitigating discriminatory biases : 차별적 편견 완화, 구조화된 데이터에 대한 머신러닝에 관한 많은 논문은 의사결정 시스템의 불공정을 강조하고 있고 전문가들이 원인을 탐구하고 해결하도록 노력해야함.
  - Debugging biases and other errors : 어떤 해석방법을 이용하여 OCL 분류를 적용할 것인가를 조사하여 특정 샘플에 대한 분류기의 성능 저하 또는 불공정성의 원인을 찾도록 해야함.

SUMMARY AND BROADER CHALLENGES AROUND OCL RESEARCH

- Bias
  - (표5 설문조사를 통해 확인된 기술적 편향)
  - 의미론적 속성 및 맥락적 속성 측면에서 정의되지 않은 온라인 상충 언어 등을 설명하는데 기술적인 어려움으로 종종 발생한다.
- Technical challenge
  - 대부분의 문제는 잘못 정의된 요구 사항에 대한 질문이다.
  - 개발전에 시스템의 요구 사항을 잘 식별하고 이러한 요구사항을 테스트하기 위한 방법을 개발하면 이러한 문제를 예측하고 수정할 수 있다.
- Adjacent challenge
  - 이 논문은 OCL에 초점을 맞춘 만큼 이미지 및 밈과 같은 다른 유형의 웹 콘텐츠에 대해 조정이 필요함.
- Structural challenge
  - 설문조사를 통해 확인된 많은 기술적, 맥락적, 의미론적 문제는 OCL에 대한 연구 및 개발이 구조화된 방식에서 근본적 원인을 찾아야한다.

Table 5. Summary of Biases Introduced in the Online Conflictual Language Detection Systems through the Design of the Data Collection Pipelines and of the Classification Models

Data Collection	**Sample retrieval**	Source & time → contextual bias; Keyword and rank biases; Topic & language biases; Representation bias; Collection of context information
	Dataset processing	Data augmentation bias; Pre-processing biases
	Dataset splitting	Information leakage → Evaluation bias
	Sample annotation	Annotator *OCL* knowledge; Annotator background; Annotation instruction; Presentation of context; Annotation aggregation
Model	**Feature engineering**	Measurement bias (context, psychology); Discriminatory features
	Classification algorithms	Aggregation bias; Discrimination bias
	Performance evaluation	Evaluation bias; Data representativeness; Metric relevance

QnA