Decoding Disbelief:

Using Natural Language Processing's Sentiment Analysis to Assess 24 Years of Unfounded Reports

Rachel E. Lovell, PhD, Cleveland State University (USA)

Lacey Caporale, PhD, Cleveland State University (USA)
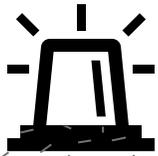
Jiaxin Du, PhD, Grand Valley State University (USA)

CSU | Levin College of Public Affairs and Education
The Criminology Research Center

CSU
Cleveland State University

CLEVELAND STATE
60
1964–2024

Funding

Collaboration

CSU | Levin College of Public Affairs and Education

The Criminology Research Center

Unfounded Classification

False Report

A clear and credible admission by the victim or strong evidence to suggest they are lying (Kelly et al., 2005)

False **OR** Baseless
An unfounded crime is one where an investigation was completed, and it was determined that a crime did not occur (FBI 2013, p. 111)

Historically misused by PDs to "hide" violent crime; myth that rape victims frequently lie

Baseless Report

The victim's account does not meet the legal definition of the crime (Lisak et al., 2010)

Purpose: What is the nature of the narratives of <u>unfounded rape reports</u> using natural language processing

DATA

- **5,638** reports of rape, Cleveland Division of Police (CDP), from population of 6,071—**all had associated sexual assault kit, 1992-2015 (24 years)**

- Reports analyzed using **Natural Language Processing's** (NLP), **Sentiment Analysis**

Capitalize on such a large sample of text of police reports

REPORTS TYPICALLY INCLUDE:

- Incident report

- Summary of the investigative activity on the case, as noted by the investigator

Overlap between these

NLP

- If going to find explicitly disbelieving language about victims, should be in unfounded and those where victims doubt/lied. ***How are these reports written different?***

- Specifically examined
  - **Unfounded** – provided and denoted by police in report
  - **Victim lied/doubted** – provided by researchers after reading report

Methods

Natural Language Processing: *Sentiment analysis for tone (polarity & subjectivity) + word count*

- Used TextBlob
- ML models possess learnable parameters (not fixed); expert knowledge is a prerequisite for rule creation (Bird et al., 2009)
- Five-fold cross-validation (Ojala & Garriga, 2010)
- Best fitting model was presented
- Trigrams were presented in two ways: 1) "raw results" 2) "grouped trigrams"
- Sentiment and subjectivity analyses were employed

CSU | Levin College of Public Affairs and Education

The Criminology Research Center

Hypotheses

RQ 1: Unfounded and victim-lied/doubted rape narratives have **not significantly changed** in tone (polarity and subjectivity) and word counts **over the observational period**.

RQ 2: Tone in *unfounded* rape narratives is statistically **more negatively opinionated, less subjective,** and the narratives have **fewer words** than other reports.

RQ 3: Tone in narratives explicitly mentioning *victims lied or were doubted* is statistically **more negatively opinionated, less subjective,** and the narratives have **fewer words** than other reports.

Key contributions
- Capitalize on such a large sample of text of police reports

- Access to nature, how are they different?

CSU | Levin College of Public Affairs and Education

The Criminology Research Center

Descriptives of the rapes, suspects, & victims

Victims
- **Female (~95%)**
- **24 years of age on average**
- **65% Black/African American, 32.6% White, 2.5% Hispanic**
- **Half of cases occurred in the 2000s, 41% in the 1990s and 6% in the 2010s**

Suspects
- **Male (99.1%)**
- **77% Black/African American, 19% White; 4% Hispanic. 2nd suspect: 76% Black/African, 21.1% White; 3% Hispanic**
- **55% had <u>fully</u> named a suspect**
- **21% arrested**

Case outcomes (with an investigation)
- **26% stalled before being forwarded to prosecution, 74% did not stall**
- **Of those that did not stall, 60% forwarded for prosecutorial review, 40% not forwarded for review**
- **Overall, 73% did not proceed to prosecution, 27% did**

Unfounded: 7% (n = 386)
Victims lied/doubted: 3% (n = 158)
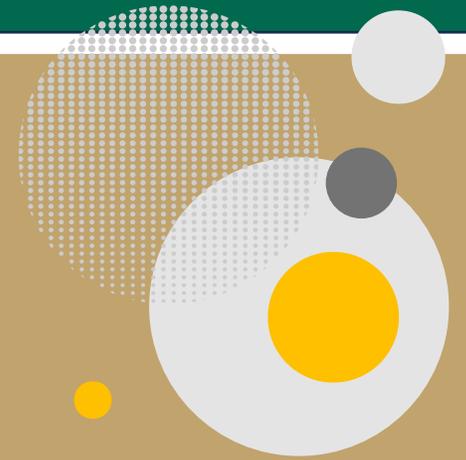
Descriptives of the Textual Data

	Statistics
# of words in analytic sample of reports:	3,931,481
# of sentences in reports	281,600
Number of word processing pages	9,157
Mean number of words in the incident report:	415
On average, every report has ... sentences:	50
# of unique words in reports:	32744
lexical diversity:	0.008

Very formulaic and procedural

CSU | Levin College of Public Affairs and Education

The Criminology Research Center

Findings for Hypothesis 1: Sentiment Analysis Over Time
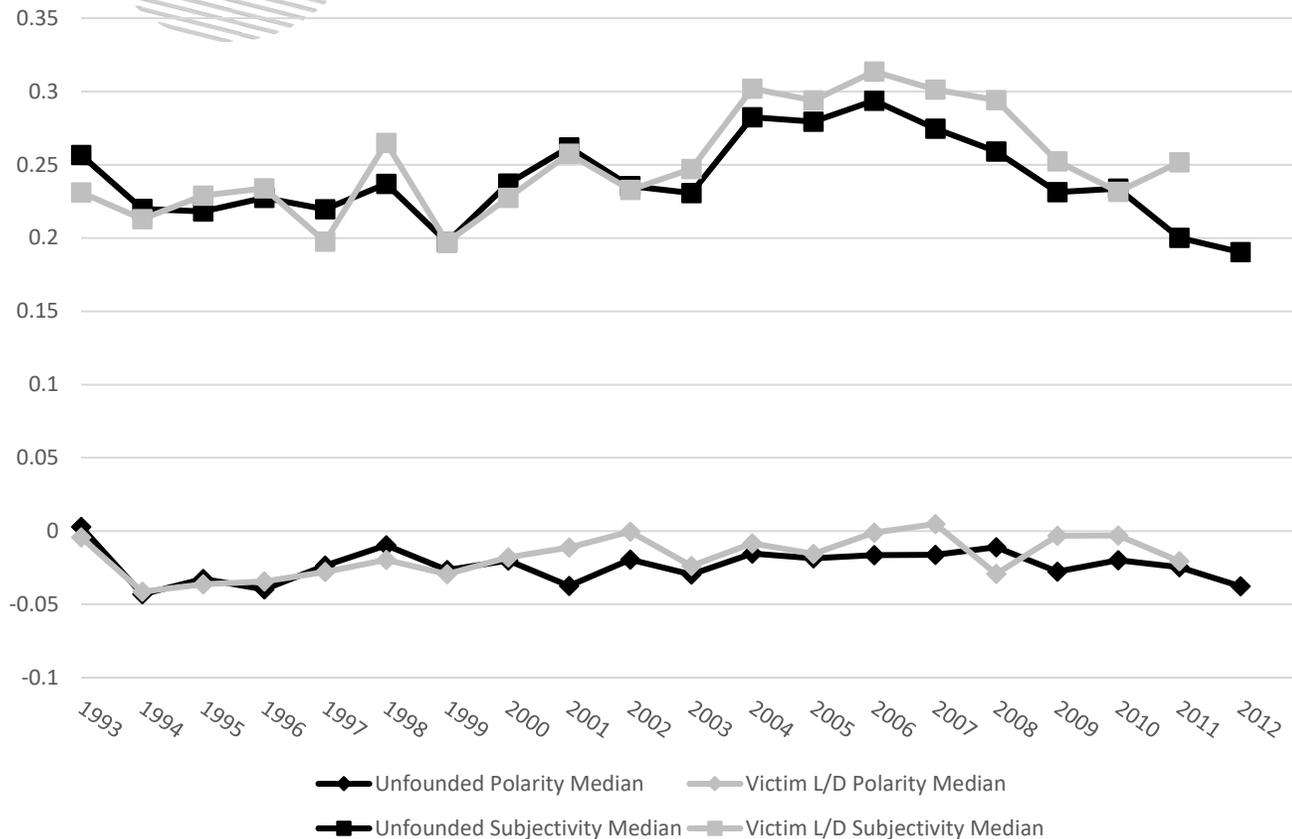
Sentiment Analysis Measures by Year

Word Counts (not shown)
- Nonsignificant over time

Polarity & Subjectivity
- More variation in subjectivity
- **All remained stable over years, even decades**
  - Small change around 2004

Findings: Hypotheses 2 & 3

Unfounded
- **Lower word counts** (*M* = 373.27, *SD* = 220.49; *p* =.002, *d* = .167) versus non-unfounded cases (*M* = 414.48, *SD* = 249.00, *p* = .002; *d* = .167)
- **More negative polarity** (*M* = -.020, *SD* = .034; *p* = <.001; *d* = .229) versus non-unfounded cases (*M* = -.011, *SD* = .037; *p* = <.001; *d* = .229)
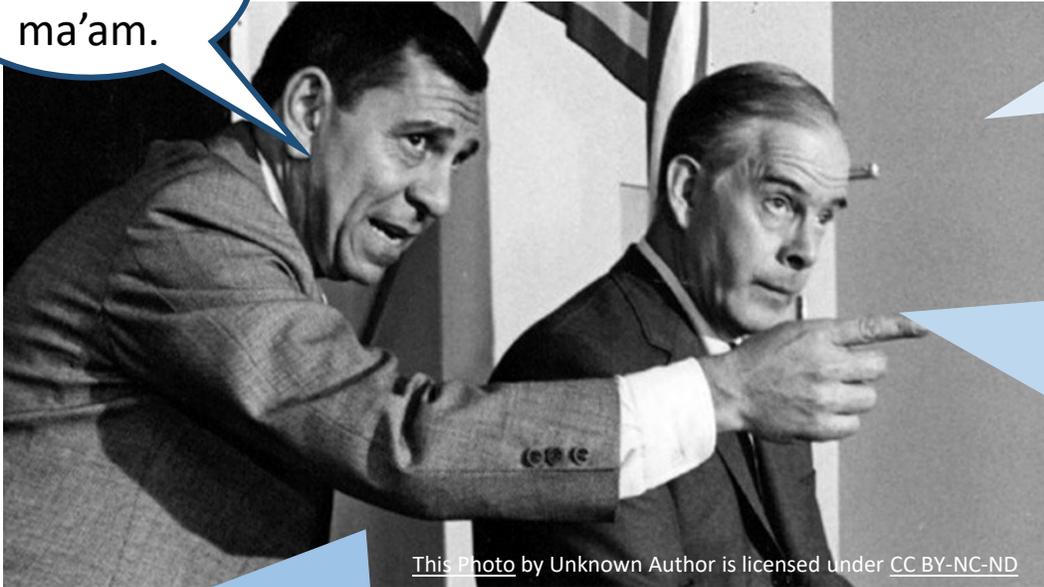- **Subjectivity** = nonsignificant

Victim lied/doubted
- **Word counts, polarity, and subjectivity** = nonsignificant

- <u>Unfounded</u> cases (false and baseless) shorter and more negatively worded, but about what and who?
- <u>Victim lied/doubted (false) reports</u> did not differ in tone (polarity or subjectivity) or length (word count), but shouldn't these be the most negative?

Follow up study: Need to know more about the actual text of the reports to understand the findings.

CSU | Levin College of Public Affairs and Education

The Criminology Research Center

Signal: Just the facts, ma'am

Just the facts, ma'am.

This Photo by Unknown Author is licensed under CC BY-NC-ND

Counter expectation
- Sentiment analysis picking up something, but it's not as expected.

Most successful cases
- Less negative words (vs. what the victim didn't do, know, say)
- More subjective words (victim's perspective)
- Those that are overall more positive tone/opinions (victim's perspective, actions)

Least successful cases
- More negative words
- No or negative subjective words
- Overall more negative tone/opinions

CSU | Levin College of Public Affairs and Education
The Criminology Research Center

Recommendations for Future Practice

Rape-specific lexicon developed, being populated

Need for lexicon specific to crimes/rape

More training for new & reporting officers

Bureaucratic burden on victim

Technological advancements: semi-structured "Grammarly" and pdf-to-text dynamic interface

Better writing, (potential) = better evidence, "thoughts" and interactions with victims

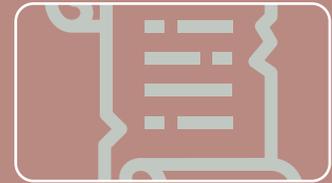**Technological advancements:** semi-structured police reports vs open narratives

CSU Levin College of Public Affairs and Education
The Criminology Research Center

Limitations
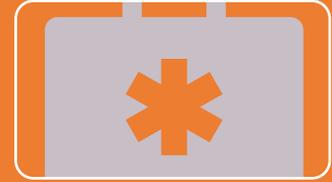
- Only rapes with kits
- Only one jurisdiction
- Not most current rape reports
- Based on official documentation

CSU | Levin College of Public Affairs and Education

The Criminology Research Center

Reference

Lovell, R. E., Caporale, L., & Du, J. (2025). Decoding disbelief: Using natural language processing's sentiment analysis to assess 24 years of unfounded rape report narratives. *Behavioral Sciences & Law: 1-16.*

CSU | Levin College of Public Affairs and Education

The Criminology Research Center

Thank you

Rachel Lovell, PhD

Associate Professor

Director, Criminology Research Center

Department of Criminology & Sociology

Cleveland State University

r.e.lovell.@csuohio.edu

CSU | Levin College of Public Affairs and Education

The Criminology Research Center