Sexual Assault Media Coverage in Israel: A Digital Humanities Experiment

Tom Kremer

Digital Humanities Tutorial, Spring 2020

Prof. Ostrow

I. Preface

My submission is a hybrid of a paper and a reflection. I wish I could deliver a standalone public-facing piece that is more friendly than a long PDF, but in light of the results and timing, I did not (more on that later). I wrote this piece in the same way that I'd explain it to someone I'd want to show it to, so it could serve as a basis for a better-formatted public-facing version. Since there is no other work product, I did not restrict myself with the assigned word count, so I apologize if that is too much, although I wanted to document everything I did.

Except for this paper, all other materials for my submission (basically, the code and it supporting files) are on the repository: https://github.com/tomkreker/SexualAssaultCoverage/. The main notebook is called *kulan*, and its structure mirrors that of this paper (except for a small reordering of the topic modeling section).

II. Introduction

In the past several months, Israel has been going through an epidemic of severe sexual assault cases, domestic violence incidents associated with the long-lasting lockdown measures, and absurdly lenient verdicts for convicted sexual offenders. While there surely have been many such cases earlier, public awareness, reaction, and media coverage in relation to recents events has been heightened.

Kulan (the feminine version of the word 'everyone' in Hebrew), a feminist activist group, has recently launced an online campign named Fair Coverage, aiming to eradicate prevalent reporting practices that perpetuate 'rape culture'. Such practices include victim-blaming, disproportionately highlighting the perpetrator's lawyers, and focusing attention on the consequences on the offenders career or life trajectory while downplaying the victim's voice and harm. A key part of the campaign is devoted to calling out media articles from online news platforms. So far, this happened sporadically and reactively, partly due to the organization's limited resources. This project is intended to support Kulan's endeavor by applying NLP, text analysis, and machine learning techniques to evaluate sexual assault media coverage in Israeli media outlets.[1] For scoping purposes, I looked at articles from four years on Walla! News.

I set out to investigate on the five following practices, captured by Kulan's Fair Coverage principles (Kulan, 2020):

1. Don't talk about what the victim was wearing and drinking or her habits.
2. Don't use 'soft' vocabulary like 'intercourse' when there are evidence for rape.
3. Don't highlight the offender's lawyer's responses.
4. Present the victim's voice at least as much as the suspect's.
5. Avoid positively describing the suspects, their credentials, or achievements.

Alas, I was not successful in effectively exploring all of these, primarily guided by the free tools I was able to access to process large amounts of Hebtew text. NLP capabilities are more limited in Hebrew, which differs from English in semantic structures, alphabet, and basically everything.

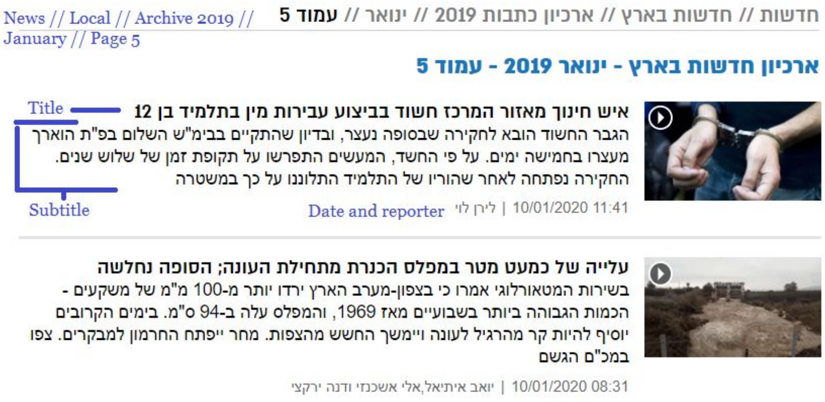
[1] Throughout this project, I use 'sexual assault cases' as an umbrella term for sexual offenses, including harassment, abuse, rape, collecting and spreading sexual content, and pedophilia, for convenience. This is what I would use in Hebrew if I were to translate, but it is possible that there is a better choice.

This report walks through what I *was* able to achieve, and ends with a (fairly long) note on potential extensions. The next section explains how the dataset used in this analysis was formed, including the various choices I had to make along the way. I then summarize findings from high-level exploration of the data via a simple word cloud and topic modeling that power a simple visualization of the dataset. This is followed by a qualitative analysis of lawyer mentions in article titles, corresponding to Rule #3, as well as a basic quantitative exploration of the verbs and adjectives used to describe men and women, as a proxy to explore Rules #1, #4, and #5[2]. I end with the promised 'if only I had an extra semester' note.

III. Dataset Construction

The dataset is made of news articles scraped from Walla News! (Walla), the second most visited news website in Israel, according to Wikipedia. It is an exclusively digital, free-of-charge media outlet which covers a wide range of topics including internal and international news, sports, and culture. I chose to build the dataset from Walla because its archive was the most organized of all three news outlets I considered (Ynet, Haaretz), hence the easiest to scrape. Also, its titles and subtitles were the most detailed, which made it easier to filter articles based on title content, as explained below.

First, I scraped the title, subtitle, date, reporter, and link to the article for all articles under the local news category. I started with January 2020 (n=304), then 2019 (n=4210), and finally for 2016-2019 (n=18586), which I used for the rest of this project. Figure 1 shows an example of the head of one scraped page.



חדשות // חדשות בארץ // ארכיון כתבות 2019 // ינואר // עמוד 5

ארכיון חדשות בארץ - ינואר 2019 - עמוד 5

איש חינוך מאזור המרכז חשוד בביצוע עבירות מין בתלמיד בן 12
הגבר החשוד הובא לחקירה שבסופה נעצר, ובדיון שהתקיים בבימ"ש השלום בפ"ת הוארך מעצרו בחמישה ימים. על פי החשד, המעשים התפרשו על תקופת זמן של שלוש שנים. החקירה נפתחה לאחר שהוריו של התלמיד התלוננו על כך במשטרה

10/01/2020 11:41 | לירן לוי

עלייה של כמעט מטר במפלס הכנרת מתחילת העונה; הסופה נחלשה
בשירות המטאורלוגי אמרו כי בצפון-מערב הארץ ירדו יותר מ-100 מ"מ של משקעים - הכמות הגבוהה ביותר בשבועיים מאז 1969, והמפלס עלה ב-94 ס"מ. בימים הקרובים יוסיף להיות קר מהרגיל לעונה ויימשך החשש מהצפות. מחר ייפתח החרמון למבקרים. צפו במכ"ם הגשם

10/01/2020 08:31 | יואב איתיאל,אלי אשכנזי ודנה ירקצי

Figure 1. The upper part of page 5 of the January 2019 article archive, Walla! News.

[2] I found Rule #2 too challenging to evaluate without careful examination of case details.

The next step was to identify the articles that cover sexual assault cases. To do this, I created a list of sexual-assault related terms, which included variations of sex/sexual, rape (in noun and verb form), the Hebrew terms for sexual harrassment, assault, and misconduct, intimate/acy, pedophile, and 'sleep with'. To account for different variations of these terms across tenses and with prepositions (which are attached to the word in Hebrew), I normalized these words as well as the joint title and subtitle text of all articles using an API-based service called HebrewNLP, which is free-to-use upon registration. The normalization engine converts, for example, '(male/female) kid', 'the (m/f) kid' 'the kids', and 'their kids', into 'kid', the masculine version (which is always shorter in Hebrew. Yep.), as it is always shorter (and in most cases a substring of the female version, in which a single letter is added). In my initial experiment with January 2020, there were 25 results. With the 4210 normalized title-subtitle texts from 2019, 348 contained one of the normalized terms.

I then manually examined the selected articles. Starting with only one month of data allowed me to also go through all articles and check whether my filtering left anything out. I found none, but there was a significant amount of false positives. It seemed that two words, 'relationship' and 'harassment', were responsible for many of them. This made sense, since these can be used in many different contexts (after normalization, relationship is equivalent to 'relative'). For 2019, I also have not left anything out, but these words created even more noise. I read the titles of all these articles, and found that they all contained one of the other terms (mostly, sex/ual), so I removed them when working with the four-year dataset. Filtering all 2016-2019 articles with the final words left 1295 articles. I then scraped the full content of these articles and normalized it with HebrewNLP to get a clean version of each article.

Even after removing the 'noisy' terms, there were some unrelated articles left in the final 1295, which is reasonable given my word-based filtering. I reviewed all titles and subtitles of the 1295 to remove false positives. More significantly than noise, which was easy to remove, it became clear that sexual assault coverage comes in many forms, and I had to decide what stories would be included. The following were the main considerations and decision points I identified, the choices I made, and the motivations behind them:

1. *High Profile Cases*: two types of sexual assault cases stood out. The first were *high profile* cases — those involving celebrity perpetrators, or ones that received repeating

media attention for various reasons. The main characteristic of high profile cases is higher-than-usual coverage. Some cases were easy to classify, but I had to set a cutoff. I included all cases with at least five articles covering them, an arbitrary cutoff which left around 3-4 named offenders out. The coverage of the more sensationalized high-profile cases is arguably different than those of one-off events (for example, the former president of Israel, convicted of rape, had numerous articles around his parole request), but both are relevant for the understanding of sexual assault coverage. I thus included articles of both kinds in the dataset, while adding a column that identifies whether or not each article covers a high-profile case. This allowed me to split the dataset at any stage of the analysis and compare the two groups. Since some articles may mention one's name but not primarily cover their cases, an article was classified as high-profile if it included one of the high-profile names identified as mentioned as above in the title or subtitle of the article. Reviewing articles which included names in the body but not the title confirmed these were indeed not covering their cases, and that articles that had the name in the title but not the body were still relevant (for example, a title or rank were used). 227 articles were classified as covering high-profile cases, and the names and descriptions of these cases are included in the appendix.

2. *Statistics, Trends, and Commentary:* I excluded 35 articles that provided statistics about sexual assault cases, reporting, trial results, or commentary about sexual assault trends. These are certainly an important piece of sexual assault media coverage, but they are qualitatively different than coverage of specific cases. To maintain a fairly consistent dataset, I decided to exclude these, although they merit their own spotlight (and probably encouragement to increase those from 35 stories in four years).

3. *Violence/Murder Cases*: I left out cases involving rape and murder. There were not many such stories, and they involve a different type of offense which might result in different coverage. One such famous case I excluded was the coverage of parole requests made by Jonathan Halo, who killed the man who raped him. I will also note that the data also does not cover domestic violence cases, unless sexual abuse also took place. While domestic abuse shares some similarities with the sexual assault domain and merits its own analysis, for dataset consistency, it was not covered.

4. *Uncommon/Borderline Cases:* most cases involved male offenders and female victims.[3] But not always. The following cases were included even though the nature of the offence or the gender dynamics were less common. First, two high-profile case: the extradiction negotiations of an Israeli female sex offender who faces allegations in Australia, and a case in which a male head of Israel's Lawyer Chamber was were interrogated for providing and receiving benefits for judges and apprentices with which he had sexual relationships. A female judge was also interrogated in this case, and covered in the data. Other less-common cases involved teachers, police officers, and prison guards who had consensual sexual relationships with students, subordinates, and prisoners (respectively, where teachers were both male and female and guards were female). Arguably, some of these differ from others (e.g. teachers and minors compared to two adult police officers), but for the sake of simplicity, and since all these were fairly rare, they were also included. I also included cases in which there were female offenders and male victims, or same-gender ones, which were much less frequent (as expected). These might affect an analysis that assumes masculine verbs and adjectives relate to perpetrators and feminine to victims, but again, there were few of these cases. [4]

At the end of this process, 947 formed the final dataset, of which 227 were high profile cases. Each row describes an article, including the date published, title, subtitle, reporter, URL on Walla, raw content, normalized title and subtitle text, normalized content, and a binary identified for high-profile cases.

IV. High-level Exploration: Word Clouds and Topic Modeling

As an exploratory entry point to the data, I built a word cloud out of the normalized content of all articles, in which the size of the word is associated with its prevalence (Figure 2). Most words revolve around the criminal process, alongside words describing the crimes: *suspect, allegations, child, arrest, sex/ual , police, investigation, violation, criminal charges, trial, defendant, criminal record*. It is not surprising that the press coverage of sexual assault cases

[3] I use 'victim' to also capture 'complainant', and because that seems to be frequently used by Kulan, whereas an alternative like 'survivor' is not commonly used for this type of cases.

[4] One other common type of cases with unordinary gender 'roles' is children victims, which had more male victims than the proportion of adult male victims. All the choices in the Uncommon Cases section were more sensitive, and could be discussed and decided upon with more nuance. Ideally, I would map the cases covered in this analysis in terms of gender 'roles', adult/child victim division, exact offense, and 'outcome'. I would then have exact counts rather than vague 'few'/'rare'. I did not do so because of time.

highlights these issues, as cases which are reported, investigated, and lead to charges, which are a small percentage of all actual incidents, get media attention.

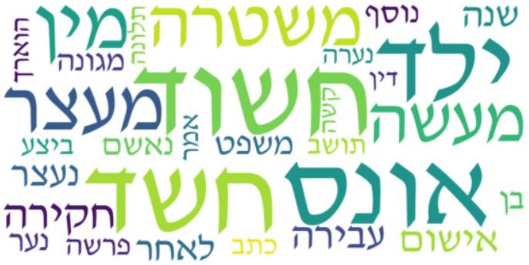


Figure 2. Word Cloud from all titles and subtitles. Main words appear in the text above, italicized.

This word cloud used a general-purpose collection of Hebrew stopwords (words to be ignored by the counter), but these criminal and procedural terms, which are clearly dominant in this dataset, are not very interesting. I thus expanded the stopwords to include hundreds of additional terms (including words that would normally be in English stopwords but were not present in my initial list) that would take out some of the noise unique to this data. This mostly included words about the investigation, trial, and punishment. as well as words describing the offenses themselves (i.e. rape, assault, misconduct, harrassment, which are obviously frequent)[5].

I then created new word clouds to the fully-filtered article content (bottom part of Figure 3), while also separating word clouds for high-profile cases and 'standard coverage' ones (left and right, Figure 3). With the standard stopwords (upper part), there is little difference between the two: only the order of the leading words *charge, police, complaint, crime, sex,* and *investigation* changes. With the more complete stopwords, high-profile cases were characterized with words that reflect the domains in which they took place, mostly the public service (municipality, local authorities), the police, and the army: *role, office, municipality, a high police rank, unit, high-in-command*. Standard coverage cases also reflect frequent domains of offenses, as well as descriptions of the assaults: *teacher, touched, video, online, attacked, taken advantage of, education, body, naked*.

[5] As I briefly mentioned earlier, assessing if appropriate terms were used for a given offense is one of the guidelines, but I could not think of a systematic way to do that. Counting the frequency of different terms, or when they come together, but without really diving into the case, I didn't find a good way to do that.

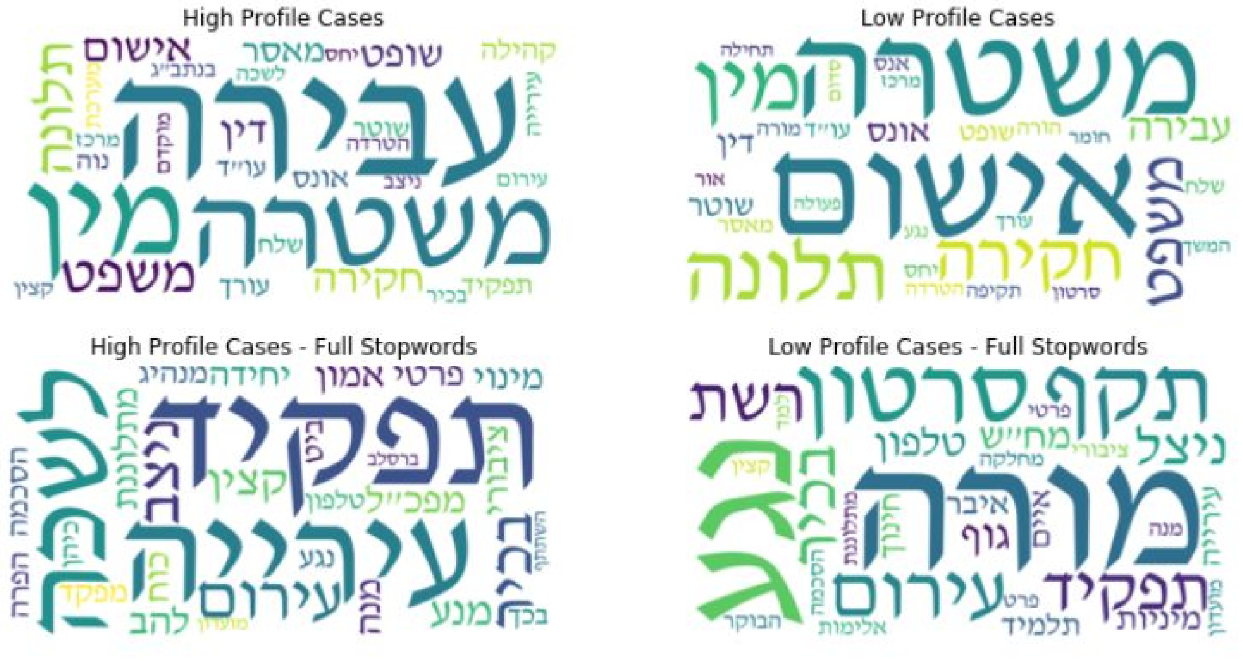


Figure 3. Word clouds constructed using partial (top) or full (bottom) stopwords and for high-profile (left) or standard coverage cases (right). Top words italicized in the text above.

Another attempt I made at an exploratory analysis was via a machine learning technique called Latent Dirichlet Allocation, which produces a 'Topic Model' for a corpus of text (the content of all articles). This technique assumes that underlying each document in the corpus are a predefined number of topics, each of which has an array of words associated with it with different degrees of prevalence. A document is represented as a mixture of words drawn from the underlying topics, as if whenever someone writes an article, for each word, they randomly draw a topic based on the mixture of topics in that document, and then randomly draw a word for that topic based on the mixture of words in that topic.

The learning process with LDAs tries to find the most likely grouping of words into a user-defined number of topics based on the way words appear in each document. These models are easier to fit when the underlying data is diverse, for example, all news articles from a given media outlet, which makes it easier to find distinct clusters of words (topics). However, there is no metric by which this model can be optimized to learn the 'best' topics — one can only look at the results and assess what each topic means and whether it makes sense. If it doesn't, they can tweek the learning parameters, the number of topics, or the stopwords, clearing out noise such that the model could find stronger groupings.

I found that five topics is the sweet-spot to create the most coherent topics, although they were still not entirely so when considering the full list of words associated with each topic. My most successful model generated the following topics, characterized by the top ten most 'relevant' words per topic, which considers in-topic frequency and distinctness compared to others:

1. **Internet/Social Media Offenses**: *video, phone, threat, car, internet, took-picture, touched, genitals, camera, intimate.*
2. **Assault Descriptions**: *club, violence, night, assaulted, consent, ran away, force, naked, body, taken advantage of.*
3. **Police/Army**: *high-ranked, position, the police department that investigates crimes by police officers, officer, a name of a high police rank, commander, another internal investigations police unit, rank, unit, served.*
4. **Public Authorities**: *municipality, local, mayor, public, prevention, office, responsible, refused, trust, position.*
5. **School**: *teacher, student, education, parent, taught, learning, touched, genitals, assaulted, youth.*

The model picked up on the common types of assault cases that appeared in Israel in recent years. There have been more than five cases of high-ranked police officers who have been charged with sexual assault allegations, as well as one high-ranked army officer. The similarity in terms in the army and police is likely responsible for their mutual grouping. Likewise, there has been a series of allegations against mayors and heads of local authorities. Another type of offense occurs 'online', including numerous operations to capture online pedophiles and arrests of people who filmed and/or spread sexual images and videos via social media. The final group of words is the actions involved in the acts themselves, after the extremely frequent ones (raped, assaulted, harassed) were removed. However, this group is also more noisy after the top terms, containing ones that would fit better in one of the other groups, so this model is very imperfect.

With this division in place, the model calculates the prevalence of each topic per document based on its words. One thing this model is useful for is to summarize articles with fairly little information (the prevalence of the five topics), and characterize them based on their dominant topic. To plot the articles by their dominant topic, we can further reduce the five-dimensional representation of each document learned by the model into two dimensions using a technique called Principal Component Analysis (PCA), which essentially tries to find the two-dimensional representation of the data that will preserve the most variance possible from the original five dimensions. Each article is then colored by its dominant topic, and X markers indicate high-profile cases. Articles appearing close to one another here have similar topic mixtures.

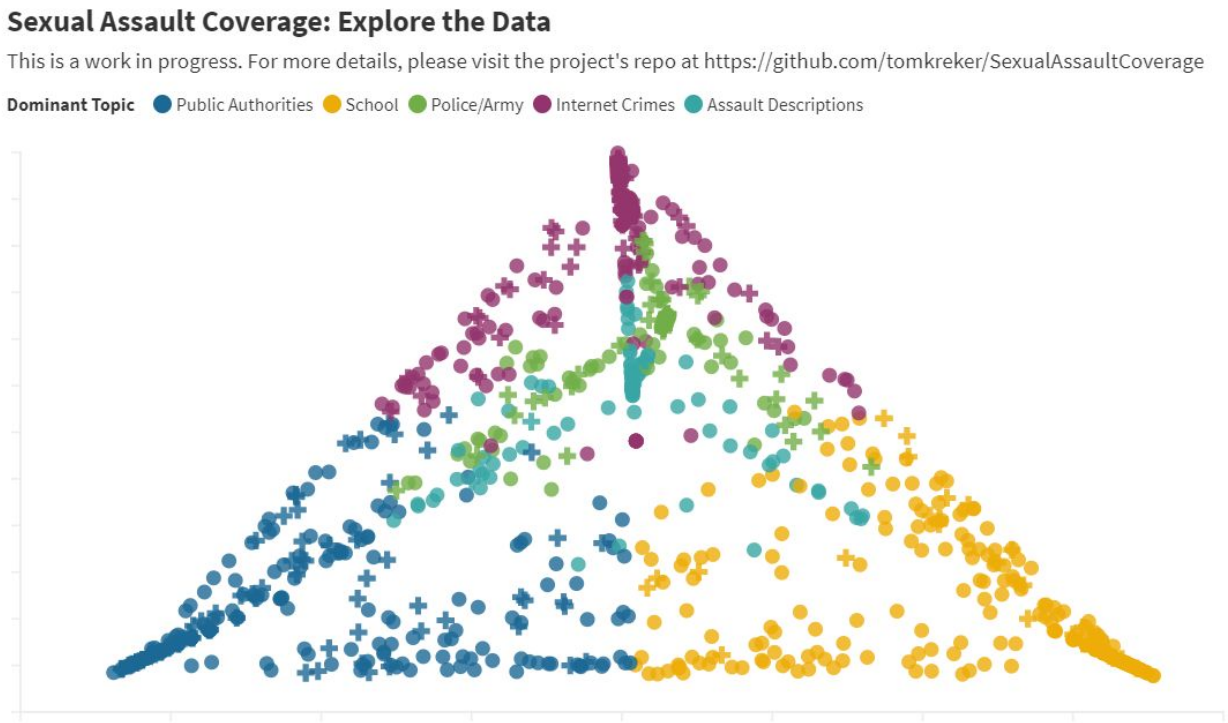


Figure 4. All articles in the dataset plotted based on their topic mixtures (axes) and colored by dominant topic. Xs are high-profile cases. Created using Flourish.

In the live version of this plot, in which hovering over a point shows the date and title of an article, enabling an exploration of the data and an on-the-fly assessment of the dominant topic classification. It affirmed that the topic model is just an approximation — while it's easy to see why some articles have their dominant topic as classified, it does not make sense for others. That's because some words in all topics are overlapping or incoherent, which is somewhat expected with relatively few articles that cover the same type of news. However, some features of the overall arrangement do make sense — Assault Descriptions are at the center, which we might expect since all documents have some of it, and it would represent ones which less clearly fall into one of the domains. Similarly, the Policy/Army group can also be expected to be central, as even after cleaning the heavy investigation-related terms, those that were left probably also appear elsewhere. Finally, it makes sense that most high-profile cases (Xs) are Public Authorities (73), Internet Crimes (68), and Police/Army (56), because that's where high profile cases happen. So even though the model is just an approximation for the actual topics, its products serve as a one-stop place to look around the data and read specific articles.

V. Lawyers

Guideline #3 is dedicated to the room lawyers get in media coverage, specifically whose lawyer's comments are being featured and what types of comments they make. While some lawyers simply say that they will review the evidence or that their client denies the allegations (even if untruthfully), others explicitly turn the spotlight on the victims, questioning their stories, integrity, and motives. Since usually the victim and offender will not directly talk to the media, their legal teams are arguably the primary means through which their voice is heard. Thus, these communications, especially ones highlighted in the titles, could contribute to the doubt, shame, and blame cast on complainants and further prohibits victims from coming forward.

In this section I set out to find whose lawyer is quoted, and what types of comments are shown. Ideally, I would want to cover all comments made by lawyers throughout the full texts, but this proved to be difficult with the tools I had — I could not reliably extract sections of text based on the mention of the different words signifying 'lawyer', as the HebrewNLP engine is not (as far as I could gather) sensitive to part-of-sentence (e.g. who is the subject). Thus, I focused on comments made in titles and subtitles, which are shorter and easier to review. One advantage of using titles is that the type of comments featured in them is in itself telling of the coverage of Walla and whose voice they highlight, since these are the parts that get the most exposure.

I filtered all titles and found 151 references to lawyers out of the 947 articles. 52 of them were irrelevant, for example, when the suspects were themselves lawyers. All others were mostly quotes of lawyers responding to allegations or a development in the investigation. I manually classified the remaining 99 into one of three categories:

1. Victim: When the victim's lawyer was being quoted, exclusively or aside the accused.
2. Defendant-Neutral: When only the defendant's lawyer was quoted, and his comment was a simple rejection of the allegation, or on the legal proceeding itself.
3. Defendant-Bad: when only the defendant's lawyer was quoted, and his comments were explicitly targeting the victim's credibility, motive, and actions.

Before proceeding, it feels necessary to acknowledge that lawyers will obviously often reject allegations against their clients and say things that promote their interests. The issue at hand is Walla's choice to highlight them at the titles of the articles (or at all).[6]

Only 16 articles of the 99 who had introduced the lawyer's comments in the titles showcased the victim's lawyer. Looking at these 16 cases, **12** were high-profile, which could be expected to feature both sides as part of the longer coverage process. The other cases involved children victims who were abused by multiple teenage boys and a rabbi accused in multiple abuse cases. There were, unfortunately, more cases of both kinds, in which the victim's side was not featured.

The content of these comments can be divided into three themes: the victim's version (5), the victim's physical or mental condition (3), and a development in the investigation (8). The first category include quotes like "*he took advantage of his authority and her innocence*" for a reality-show star who assaulted his dance students, "*there were multiple face-to-face harassments, not just via phone*" in the high-profile case of a high-ranked police officer, and "*a rape has occurred*" in a high-profile case in which a young British woman was raped by multiple Israeli teenagers. The latter include remarks on releasing suspects, on the criminal justice system, and on a court's decision to allow an accused officer to return to duty. All three are arguably important, but are a rare occurrence compared to the featuring of the offender's lawyers.

The other 83 titles only featured the defendant's lawyer. 53 of those had some form of denial of the allegations in a succinct way. The other 30 can be roughly divided into five main themes of particularly problematic comments. Some of them are problematic for their content, and not for their coverage, which was at least implicitly calling them out. Let's dive in.

Motives: Questioning the victim's motive

- In February 2016, a secretary in a local council pressed charges against the head of the council, claiming that he touched her against her will, and started harassing her when she refused. His lawyer responded: "*a cynical attempt to achieve her goals*".

[6] I am not knowledgeable enough to understand whether there's some inherent journalistic norm that obliges Walla to quote some comment, whatever it is, which would take some of the responsibility away from them and onto the lawyers themselves. Again, what remains is the choice to highlight them. But I am well aware that I may not be thinking about something here — I am not familiar enough with the domain.

- December 2016, a 20-years-old man was arrested for sexually abusing his mentally-disabled sister. His lawyer was quoted in the title claiming in court that "*it is a false accusation of his sister who tries to hurt her family for not marrying her because of her disability. It is the result of her envy in other girls her age"*.
- Just five days earlier, the lawyer of a rabbi accused for raping a 14-year-old girl at his school was quoted saying "*the complaint is an excuse for her losing her virginity before her marriage*".

Story: Questioning the victim' story

- After a court has convicted Yossi Cohen in March 2016 for kidnap and sexual assault after he tempted a girl to his taxi and assaulted her, his lawyer commented: "*the child's spontaneous response was that the touches were not sexual, she changed her version"*.
- October 2016, A 60-year-old man was arrested for allegedly committing assaulting a 4-year-old boy after his aunt, who filed the report, claimed she saw it happen. His lawyer was quoted saying "*my client claims the aunt is imagining*".

Victim-blaming: Highlighting the victim's behaviors

- A pub owner was charged with rape after providing alcohol and drugs to a 17-year-old girl. During the proceeding (April 2016), his lawyer showed videos of her behavior throughout the night, claiming that they show "*like a thousand witnesses*" that she acted like a "*slut*". Later in the article, they featured the long condemnation of the director of a center of a national help center for sexual assault victims. Walla also reported a complaint was filed against this lawyer, who kept justified his actions.
- Similarly, in April 2017, a lawyer of three rape suspects was quoted saying "*she was wearing a thong*". This case also led into a short public outrage, which was also covered by Walla in the article and ones that followed.
- September 2016, A musician who was arrested under the suspicion of raping a 13-year-old with three of his friends had his lawyer featured saying "*she admired him*".

Poor man: Emphasizing the wrongdoing done to the defendant

- March 2016, a lawyer 'related to' (as originally written) Former President Moshe Katzav, who was convicted with rape and sexual assault, during his parole comittee: "*Katzav and his family have experienced a crisis; the media is after him*".

- June 2016, the lawyer of a police officer convicted for offering a woman to perform oral sex on him to avoid a fine was quoted: "*he hopes the police will let him stay in office*".
- May 2019, the lawyer of a man charged for rape has said that "*the complaintant's tragedy is hard, but also what the defendant has gone though*".

Good man: The defendant is a good/successful/appreciated person

- The lawyer of Alon Kastiel, a high-profile case convicted for multiple rape and assault cases, was quoted after his arrest in December 2016 for four rape cases saying "*he treated women with respect all his life*".
- January 2018, a police officer who was suspected of forcibly kissing and covertly taking pictures of women had his lawyers quoted: "*he is an excellent police officer, the pictures were received on WhatsApp from his friends*".

Let's digest all of these. First, the few outright victim-blaming comments that stood out in the titles were 'good' examples in terms of Walla's coverage, in that they flagged these comments as worth attention, and almost always (at least in the major cases) followed with condemnation, although mostly from an external organization and not the victim or its lawyers. After reading the articles, these actually seem like positive coverage examples, even if not explicitly voicing the victim's side.

The 'story' group is a bit more difficult, and it is also more fuzzy, because essentially, all suspects who deny allegations claim, to some extent, a different story than that asserted by the victim. What striked me as more damaging in the more explicit form of version disputes communicated via lawyer responses in titles is that for someone who does not read the article, that is the 'last word' being said, and that undermines the credibility of victims coming forward, for which Kulan's repeating theme in their social media commentary is "I believe you". It is as if the title gives no room left to entertain the experience of the victim and the effort it took to come forward, before it is immediately contradicted by a lawyer's response. And this story goes to the next level in the 'motive' group, in which this alternative version is not only about what happened (whether or not touched were sexual, or whether an aunt has seen what she claims to see), but *why* they come forward and claim it. This involves a specific story about money, status, and even pure vengeance, that goes even further to undermine victims' credibility.[7]

[7] Here, like probably everywhere in this section, there's probably a lot more to say, literature and statistics to draw on, theory to bring in, and counterarguments to entertain, but again, it will be left as is.

For the last two groups, ‘poor man’ and ‘good man’, I see even less reasons to highlight them, even if the lawyers thought it would do good for their clients to say them, as emphasized continuously by Kulan and many others. The hardships of the offender (especially a convicted rapist), their past achievements, or their great personality (…) are simply not what these cases are about — they are suspected in committing horrible crimes, and to dedicate title ‘real-estate’ to discuss them further marginalized that victim’s voice and implicitly questions their credibility (by suggesting that it is unlikely that an excellent officer would do such things).

To conclude: victim-blaming must be featured only to be called out to be such and condemned, and from my title review, it normally is; alternative stories or suspected motives are more complicated, but should also not receive as much attention in titles as they did, especially without at least accompanying it with equal attention to the victim’s responses; offender hardships or personality adorations in titles (and elsewhere) should be reduced to zero.[8]

VI. Characterization of Perpetrators and victims

The ‘poor man’ stories are a good transition into another dimension of the data I hoped to explore: the way perpetrators and victims are characterized in the media reports. What are they doing? What is said about them, by who? How much do they speak? Such an analysis would benefit from a tool like ‘BookNLP’ described in Underwood (2019), which given the text of a certain novel, grouped characters by all aliases given to them and found all references made to or by each character. Unfortunately, I was not able to find an available tool to do that.

An approximation of this process, however, can take advantage of the gendered structure of Hebrew, and the fact that in the vast majority of assaults, the perpetrators are male and the victims are female.[9] I used the Analyzer functionality of the HebrewNLP library, which assigns a part-of-speech for each word in a text and its gender, to extract all singular, gendered verbs in the full texts. Beyond the reasons mentioned above, conceiving of masculine verbs as relating to perpetrators is simplifying because the police, court, prosecution, lawyers, and judges are all

[8] With a more comprehensive classification and more data (e.g., if I had counted content references too), it might be interesting to examine the changes in these dynamics over time. Also, throughout this entire section, I should acknowledge that some gap in lawyer’s responses makes sense, since the victim’s side is often given as a description of the allegations rather than a quote from a lawyer.

[9] There are exceptions, of course, particularly in cases with child victims, and removing them would have made such analysis stronger, but I did not do so. Another disadvantage is that plural verbs are not gendered, which would miss out on cases of group offenders.

gendered, and thus could obscure results. I removed verbs of both genders that are about the legal or criminal, proceedings as well as non-informative ones and others which clearly describe entities that are not persons. Leading removals included *arrested, committed, said, charged, released, claimed, complained, submitted, and filed*. Two exceptions I left were *denied* and *admitted*, which are clearly about the person and I wanted to compare. I then counted the number of verbs of each kind and plotted the top 20 verbs for each gender.

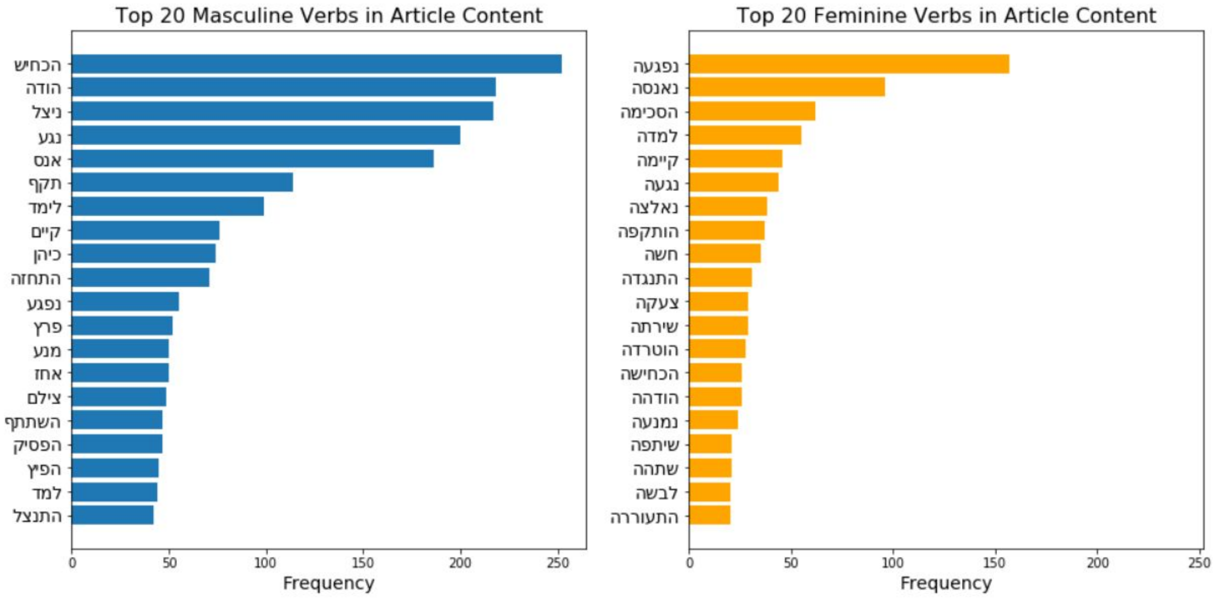


Figure 3. Top 20 masculine and feminine verbs. **Masculine**: denied, admitted, took advantage of, touched, raped, assaulted, taught, had (intercourse), held office, impersonated, was hurt, broke (into), prevented, held, took picture/video, participated, stoppe, spread, studied, apologized. **Feminine**: was hurt, was raped, agreed, studied, had (intercourse), touched, was forced to, was attacked, felt, resisted, shouted, served, was harassed, denied, admitted, refrained, shared, drank, worn, woke up.

The results are hardly surprising. Men are primarily performing the acts of rape, assault, and harassment, with some words attesting to common domains (*studied, served*). Women are described primarily using passive language: *was raped, was assaulted, was harassed*. Less frequently their actions during the events are described actively (*resisted, shouted*), but closing the list are words describing circumstances of drinking, as well as references to clothes (*worn*).

Summing only the top 20 of each category, masculine verbs were present 1988 times compared to 845 feminine ones. Since these are filtered to fairly reasonably refer to the perpetrator and victim, it is one (imperfect) measure of who gets more space, alongside the predominantly passive language describing women. Looking at the full list of verbs (not just the top 20), the proportion remains similar (9140 masculine vs. 4421 feminine), although I did not filter those as for the top 20 (in which I continuously updated the stopwords until the top 20 of both categories were more-or-less clearly about people and informative.

I also tried running the same analysis separately for high-profile cases and the 'standard coverage' cases. The figures for standard coverage are very similar to the one above, which makes sense — these are most cases. For high-profile ones, however, a few more things stand out (Figure 6).

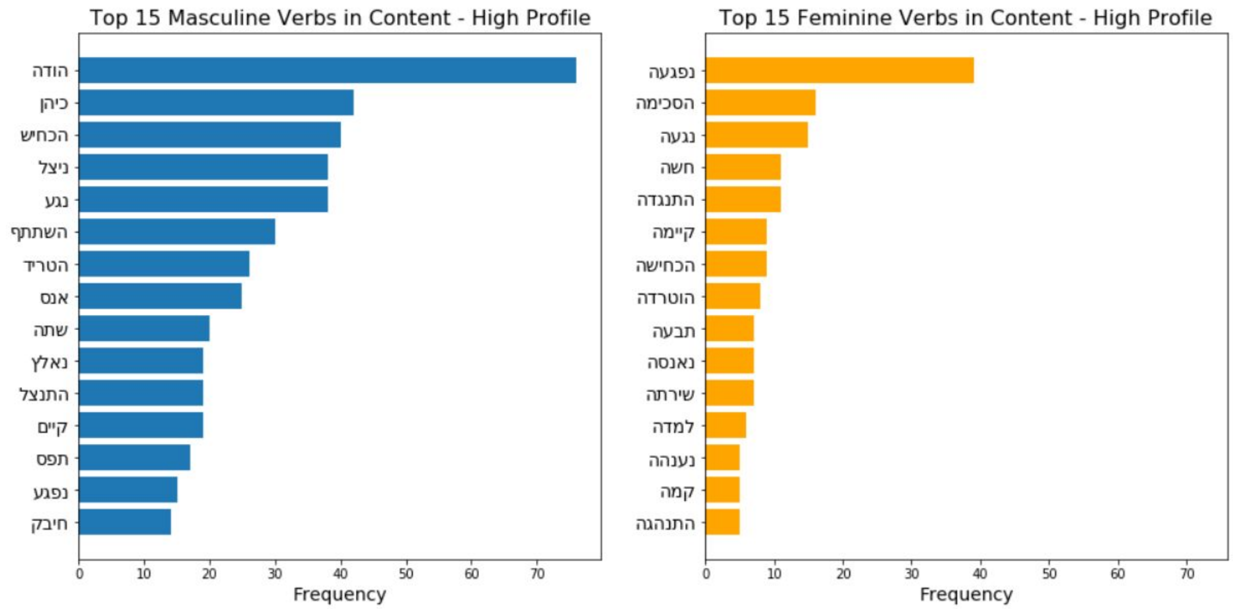


Figure 6. Top 15 masculine and feminine verbs for high-profile cases. **Masculine**: admitted, held office, denied, took advantage of, touched, participated, raped, drank, was forced, apologized, had (intercourse), caught/grabbed, was hurt, hugged, agreed. **Feminine**: was hurt, agreed, touched, felt, resisted, had (intercourse), denied, was harassed, sued, was raped, held office, studied, consented, got up, behaved.

For men, while standard cases start with admitted/denied, followed by action verbs *took advantage, touched, raped, and assaulted, here held office* is second in frequency after *admitted*, before *denial* and the same action verbs follow. While it makes sense that it will be more frequent, it might suggest that there are more references to the person's role than their actions compared to standard coverages. Furthermore, the verb *attacked/assaulted* went down to being only the 40th most frequent disappeared from the masculine list of actions, whereas for standard coverage cases it appeared about 100 times at 7th most frequent, a bit less frequently than *harassed*. All other terms (*raped, harassed, took advantage, touched*) are present at the top in both categories. It is unclear why this difference exists — perhaps its harsher connotation and the fact it is not as clearly defined as rape are responsible. The ratio of masculine (438) vs feminine (160) verbs in the top 15 ones was the same as other cases.

For another angle at the characterization of both sides, I looked at masculine and feminine adjectives. After repeating the process described above, only this time with adjectives, it became clear that it is much more difficult to derive a meaningful representation by manually filtering words; the ambiguity was stronger. Here Underwood's BookNLP would have made a bigger difference. I then took a different approach, and manually searched some adjectives that I encountered on my reading and were highlighted by Kulan in different opportunities as commonly used positive adjectives for men suspects. Two of them, *valued* and *normative*, yielded 12 and 16 minions in the full contents, respectively. In most cases, these descriptions were quotes by lawyers, the charges, or people related to the person being charged of the crimes. As such, infuriating as these descriptions may be, I am not sure about the media's role in presenting them. If I could argue for misplacing them from the title earlier, it is less clear on what grounds can they remove comments from the body, as long as there is ample enough space dedicated to the accusations and the victim's side, which was the case in all the examples above.

VII. Wishlist

Here are the five main things I would do if I had an extra semester to work on this project:

1. I would try hard to find a way to apply a tool like BookNLP to this corpus to extract the verbs, adjectives, and all references made by and of 'characters' in each article. This would enable looking at things like passive vs active verbs, typical verbs used by and describing each person, positive descriptions of accused men, and the content of lawyer's comments more comprehensively and accurately.
2. With more accurate quantitative estimation of these facets, I would make a time-series section that examines these statistics (e.g., how many lawyer references were of each side and of which of the three types I identified), in the four years of the data.
3. I would add a whole section to this paper that summarizes the statistics found in the 35 articles I found on the different domains of sexual crime or via other data sources to provide better context and backing to some of the claims I make in passing throughout.
4. I would create a full mapping of cases in this data based on the dimensions I mentioned in the footnote above. I would then either update the visualization or create a new one

that will allow anyone to explore these dimensions (starting to shift towards the sort of projects that highlight individual cases and detail that get lost in the masses — Israel is small, luckily the mass is more manageable).

5. I would dedicate more care and attention to the qualitative analysis of the types of comments made by lawyers, their implications, why whey matter, and what they reflect.
6. Last but not least, I would share this work with Kulan (hopefully after some more work, but not a lot), or at least make a public-facing (not PDF) version of this write-up, because it's a digital humanities tutorial and I am disappointed that I am submitting a PDF files, because of the value of collaboration, their domain expertise, and #dhimpact.

If I end without a conclusion, will it mean this is not the end?

Appendix: High Profile Cases

Full disclosure: after running most of the code and analysis with a set of names, I realized I should add two other ones. Most of the articles covering them were already included, so the difference would mostly be in terms of high-profile vs not, but the number of cases that would be added is fairly small. I don't think it would have changed a lot, but as I was rebuilding the dataset to finalize this paper with these inclusions, HebrewNLP crashed (I did rest my API code though, and it was working well for a while, so I hope it's not me). So the high-profile numbers in the paper do not include them. These are their names, number of cases covering them, and roles or occupation when their case was exposed.

1. Eliezer Berland (52), Rabbi
2. Moshe Katzav (46), former President
3. Alon Kastiel (21), real-estate trader
4. Ofek Buchris (20), high-ranked army commander
5. Itamar Shimoni (18), Mayor
6. Malka Leifer (18), wanted in Australia
7. Efi Nave (12), Head of Lawyer Chamber
8. Roni Ritman (10), high-ranked police officer
9. Niso Shaham (9), high-ranked police officer
10. Itzhak Cohen (8), a former judge
11. Dani Biton (6), a father of a famous singer
12. Ronen Biti (5), a father of a millennial celebrity
13. Excluded ones: the 'Cyprus' incident I mentioned in the text (11), and Mohammad Katusa (17), who was temporarily suspected in raping a 7-year-old girl and later released.

LO Appendix

- #dataconstruction — I ended up spending more time on this LO than I originally thought, which is a lesson about what's involved in building a dataset. I tried to be very explicit about the data I scraped and the choices I made when cleaning and selecting. Not all of them were rigorously motivated, but at least they were somewhat reasonable and transparent.
- #dhquestion - I was trying to build a dataset and analysis in light of the overarching goal of evaluating the stories against Kulan's guidelines. Even though my responses to the questions were limited at times, I learned a lot about what answering them involves. I tried touching upon more dimensions of these questions that I did not get to explore in the Wishlist section.
- #toolsandtechniques - Spent a lot of time figuring out the HebrewNLP, playing around to make a decent topic model and a visualization on top of that, and scarped lots of data
- #dhimpact - as I mentioned in my end note, the motivation for this project was not to have it be a PDF, so I really do want to continue working on it and sharing it, at least in some capacity. I spent a lot of time making the Jupyter Notebook readable and explaining all my steps in this paper very explicitly to make sure that someone without much domain knowledge can understand what I did and get behind it (or easily change the process), so that it could actually be used to its intended purpose and potentially be developed into a public-facing piece, cover more media outlets, turn into a visualization-centered project, or wherever Kulan's wind will take it.

References

HebrewNLP (n.d.). HebrewNLP. Retrieved from: https://hebrew-nlp.co.il/

Underwood, T. (2019). Distant horizons: Digital evidence and literary change. Chapter 4. The University of Chicago Press.