CSIREVIEW

# Gender and Whistleblowing Online

## Comparative Analysis of the Twitter Discourse around Peiter Zatko and Frances Haugen's Revelations – A look into gender-based elements

*Giovanni Maggi**

**1. Introduction -** When discussing the impact of digital technologies on social inequalities in the early days of the internet, the claim was that this technology would act as a great equaliser.[1] The ideas was that the nature of online communication would allow for a more pluralistic participation in the public sphere, which in turn would have led to flattening social inequalities.[2] However, as the internet developed, we have started to see many of the same dynamics prevalent in the offline world, play out online as well.[3] The pendulum seems now to have swung towards a more dystopian view of the online ecosystem, one in which hate speech, disinformation, and harassment is not only present and complicated to tackle, but also drive engagement on social media.[4] These online phenomena have serious implication with regard to how people participate to online discussions, leading some to self-censor in light of precedent incidents.[5] The focus of this essay is to study gender-based harassment in one specific case: that of tech-sector whistle-blowers, individuals who are of crucial importance in bringing to light information which is in the public interest. In this regard, section 1 outlines the theoretical framework within

* Indirizzo e-mail: giovanni.maggi@sciencespo.fr.

[1] See for example Yochai Benkler, *The Wealth of Networks. How Social Production Transforms Merkets and Freedom* (London: Yale University Press, 2006).

[2] Jen Schradie, 'The Great Equalizer Reproduces Inequality: How the Digital Divide Is a Class Power Divide', in *Rethinking Class and Social Difference*, ed. Barry Eidlin and Michael A. McCarthy, vol. 37, Political Power and Social Theory (Emerald Publishing Limited, 2020), 81–101, https://doi.org/10.1108/S0198-871920200000037005.

[3] See for example danah boyd, *It's Complicated. The Social Lives of Networked Teens* (London: Yale University Press, 2014).

[4] This aspect is currently being tackled by tech platforms – see Josh Constine, 'Facebook Will Change Algorithm to Demote "Borderline Content" That Almost Violates Policies', *TechCrunch* (blog), 15 November 2018, https://techcrunch.com/2018/11/15/facebook-borderline-content/.

[5] Sarah Sobieraj, 'Bitch, Slut, Skank, Cunt: Patterned Resistance to Women's Visibility in Digital Publics', *Information, Communication & Society* 21, no. 11 (2 November 2018): 1700–1714, https://doi.org/10.1080/1369118X.2017.1348535; Sarah Sobieraj, 'Gender, Digital Toxicity, and Political Voice Online', in *The Oxford Handbook of Digital Media Sociology*, ed. Deana A. Rohlinger and Sarah Sobieraj (Oxford University Press, 2022), 0, https://doi.org/10.1093/oxfordhb/9780197510636.013.29.

which the analysis take place. It does so while also emphasising the wider societal implications that the detection of gender-based harassment and aggressions might have for future whistle-blowers. Moreover, section 2 focuses on the methodology employed for the study – detailing the data collection and analysis practices employed – while section 3 explores the findings. The paper finally concludes and discusses the limitations of the methodology hereby employed.

**2. Literature Review -** Democratic debate and its underlying power dynamics in our society play out through a complex web of social discourse and interactions. In the 70s, Foucault claimed that the ritual of discourse creates a micro-physics of power based on the relational interaction between two people.[6] Discourse does not exist outside of a bilateral exchange where clear power dynamics between an interrogator and its interlocutor are at play.[7] These discursive interactions create regimes of truth which bind individuals to certain practices, shaping their preferences by manipulating their epistemological condition and infosphere through their everyday exchanges and practices. In turn, this has effects on what is socially accepted, shaping individual's perceptions of what they are legitimated in doing. These theoretical intuitions are at the centre of the investigation here proposed.

We start from the claim that both our linguistic habits as well as our discursive interactions "form the limits of our reality. [They are] our means of ordering, classifying, and manipulating the world".[8] In fact, as shown by Ng and Deng[9], language is a means of reproducing and maintaining society's status quo power dynamics. As they explain, linguistic interactions are built on semantic and grammatical rules as well as manners of speaking which reflect the "historical male dominance in society".[10] These reiterate

[6] See Michel Foucault, *The History of Sexuality*, vol. 1 (New York (NY): Random House, 1978); Michel Foucault, *Power: The Essentail Works of Michel Foucault 1954-1984* (London: Penguin Classics, 2020).
[7] Foucault, *The History of Sexuality*.
[8] Dale Spender, *Man-Made Language*, 4th edition (London: Rivers Oram Press, 1980). Page 3.
[9] Sik Hung Ng and Fei Deng, 'Language and Power', in *Oxford Research Encyclopedia of Communication* (Oxford University Press, 22 August 2017), https://doi.org/10.1093/acrefore/9780190228613.013.436.
[10] Ng and Deng. p. 8.

stereotypes,[11] and shape people's perception of what is and is not acceptable in the public sphere, and, in turn, influence how democratic debate plays out. Further, as one author puts it, "there has never been equal access to mainstream publics, nor have all voices or styles of communication been valued equally when included in the conversation".[12] Secondly, and more explicitly, power dynamics are also embedded into discursive interactions themselves. In fact, the disclosure of one side's higher status, or physical power position in the conversation leads to specific linguistic patterns which often negatively impacting the weaker party.[13]

This second dynamic, especially looking at the risks of gender-based violence and harassment, restricts certain demographics' access to democratic life and debate – in turn reducing their influence and possibilities to both change linguistic and discursive habits, as well as limiting the representation of those demographics' concerns and points of view. Crucially, gender (together with race and class) is at the centre of these dynamics. As Sobieraj pointed out, "women's use of public space is shaped by the looming possibility of gender-based incidents that threaten to undermine their freedom, comfort, and safety".[14] In the offline world, for instance, women's mental and geographic understanding of a city – their mental map – is restricted by the fear of gender-based violence and harassment.[15] Similarly, Sobieraj[16] shows this to be the case also in online public spaces, where the likelihood of gender-based (micro)aggression is higher for female than for males.[17] In fact, women's impact and participation in the digital public space is limited through the use of the techniques of "intimidating, shaming, and discrediting", all of which target women's bodies, gender, but also weaponise their fears of rape or physical violence.[18] Importantly, there is a habit to employ misogynistic epithets, sexual stereotypes, as well as attacking women's appearance and sexual

[11] Camiel J. Beukeboom, 'Mechanisms of Linguistic Bias: How Words Reflect and Maintain Stereotypic Expectancies', in *Social Cognition and Communication*, Sydney Symposium of Social Psychology (New York, NY, US: Psychology Press, 2014), 313–30.
[12] Sobieraj, 'Bitch, Slut, Skank, Cunt'. p. 1702.
[13] Ng and Deng, 'Language and Power'. p. 6.
[14] Sobieraj, 'Bitch, Slut, Skank, Cunt'.
[15] Gill Valentine, 'The Geography of Women's Fear', *Area*, 21, no. 4 (1989): 385–90.
[16] Sobieraj, 'Gender, Digital Toxicity, and Political Voice Online'.
[17] Maeve Duggan, 'Online Harassment', *Pew Research Center: Internet, Science & Tech* (blog), 22 October 2014, https://www.pewresearch.org/internet/2014/10/22/online-harassment/.
[18] Sobieraj, 'Bitch, Slut, Skank, Cunt'.

behaviours with the aim of discrediting a women's claims or skills.[19] This has two effect: (i) it redirects the discourse's focus towards the gender of the individual rather than keeping it on the actual topics of discussion, and (ii) it creates precedents that play into women's mental map of where they can and cannot contribute to online discussions to prevent such (micro)aggressions – which have effects on women's psychology. In short, gender-based aggressions and commentaries "can be understood as flailing attempts to reassert the centrality of gender difference, – and the gender inequality that comes with it".[20]

These dynamics have been well-researched in the context of politically active actors in the public sphere. In fact, there are evidence showing that female politicians are more likely to be victims of gender-based harassment,[21] and the same has been found in the case of journalists.[22] Nevertheless, evidence are missing for whistle-blowers, a category of people which heavily influences democratic debate by bringing the public's attention on specific topics of discussion. These are often highly relevant to the public understanding of political events, with well-known historical examples such as Edward Snowden's revelations about the US National Security Agency's mass global surveillance, or Daniel Ellsberg's about the Vietnam War – the so-called Pentagon Paper. Whistleblowing actions have the potential to shift narratives around political and social topics, leading to policy change on the issues of the revelations.

Here two aspects which stand at the basis of this study must be emphasised. First, the effectiveness of whistle-blowers' revelations depend on the public focus on their specific issues. If the gender or identity of the actor leaking the information becomes a topic of discussion in itself, then we can expect the effectiveness of the leak to not maximise its impact. Secondly, it has been shown that whistle-blowers of different genders have

[19] Sobieraj.

[20] Sobieraj. p. 1708.

[21] Sarah Sobieraj and Shaan Merchant, '7 Gender and Race in the Digital Town Hall: Identity-Based Attacks Against US Legislators on Twitter':, in *7 Gender and Race in the Digital Town Hall: Identity-Based Attacks Against US Legislators on Twitter* (De Gruyter Open Poland, 2022), 89–110, https://doi.org/10.2478/9788366675612-008; Elena Musi et al., 'Is toxicity towards Italian politicians gendered? A multi-level analysis of hate speech on Twitter during election period', *Luiss Data Lab* (blog), 2023, https://datalab.luiss.it/ricerche/hate-speech-twitter/.

[22] Silvio Waisbord, 'Mob Censorship: Online Harassment of US Journalists in Times of Digital Hate and Populism', *Digital Journalism* 8, no. 8 (13 September 2020): 1030–46, https://doi.org/10.1080/21670811.2020.1818111.

different reasons for blowing the whistle[23] – and thus tend to leak different types of information. It is therefore desirable to have gender balance in this category – i.e., to incentivise all genders – in order to maximise the variety of information of public interest that get leaked. In this regard, the effects of gender on whistleblowing intentions are still debated in the academic research.[24] Some have found significant effects of gender,[25] other studies found positive effects of being female or male,[26] [27] while others found no effects.[28] All of these are accompanied by thorough theoretical explanations of why they may be. Nevertheless, research finds that whistle-blowing intentions are indeed linked to perceived support of the individual – which is often linked to their identity.[29] The intuition here is that similarly to the cases investigated by Sobieraj, the gender of whistle-blowers' influence on the online debate that develops around their arguments, shapes their preference and incentive to undertake certain actions. We have seen that female whistle-blowers are perceived in an unfavourable way, and that their actions are more likely to face retaliations.[30] This is often linked to the fact that "women are expected to be compliant, unassertive, and not particularly vocal; these societal roles carry over to the workplace".[31] These, together with other gender-based stereotypes, discourse, and harassment have the potential to disincentivize women undertaking whistleblowing actions in similar ways in which gender differences already impact women's voices in the online public sphere.

[23] Mary Saade, 'Women & Whistleblowing', *Hastings Journal on Gender and Law* 34, no. 1 (2023).

[24] Abhijeet K. Vadera, Ruth V. Aguilera, and Brianna B. Caza, 'Making Sense of Whistle-Blowing's Antecedents: Learning from Research on Identity and Ethics Programs', *Business Ethics Quarterly* 19, no. 4 (October 2009): 553–86, https://doi.org/10.5840/beq200919432.

[25] Alansyah Jaka Nur Adli and Nurul Hasanah Uswati Dewi, 'The Effect of Personal Cost, Anticipatory Socialization, and Gender on Whistle-Blowing Intention', *The Indonesian Accounting Review* 7, no. 2 (10 December 2017): 211, https://doi.org/10.14414/tiar.v7i2.1601.

[26] Vadera, Aguilera, and Caza, 'Making Sense of Whistle-Blowing's Antecedents'.

[27] Marcia P. Miceli and Janet P. Near, 'Individual and Situational Correlates of Whistle-Blowing', *Personnel Psychology* 41, no. 2 (1988): 267–81, https://doi.org/10.1111/j.1744-6570.1988.tb02385.x.

[28] Dian Fitria Handayani and Nayang Helmayunita, 'Women and Whistle-Blowing: Gender in Reporting Channel and Moral Reasoning to Report the Fraud in Procurement Processes in The Government Sector' (3rd International Conference on Accounting, Management and Economics 2018 (ICAME 2018), Atlantis Press, 2019), 391–400, https://doi.org/10.2991/icame-18.2019.43.

[29] Vadera, Aguilera, and Caza, 'Making Sense of Whistle-Blowing's Antecedents'.

[30] Michael T. Rehg et al., 'Antecedents and Outcomes of Retaliation against Whistleblowers: Gender Differences and Power Relationships', *Organization Science* 19, no. 2 (2008): 221–40.

[31] Saade, 'Women & Whistleblowing'. p. 48.

**3. Research Design and Methodology -** *3.1 Context and Research Question* **-** Building on the theoretical framework outlined above, In this study, we thus aim to gather evidence and investigate the presence of such elements in the online public discourse around whistleblowers. Specifically, we aim to address the lack of evidence by looking at gender-based elements, harassment, verbal attacks, sexism, and focus of the discussion in the online public discourse. The research question here under investigation is therefore:

*RQ: How does the gender of a whistle-blower change the online discourse around their revelations?*

This paper aims to answer this research question by conducting a comparative analysis of the online discourse around two recent tech-sector whistle-blowers: Frances Haugen and Peiter Zatko. Let us start with a bit of context.

Frances Haugen (female) was product manager in Facebook's civic integrity team from 2019 to 2021 – a team which deals mostly with disinformation. During her time there she lived through some of the decisions made during the highly controversial 2020 US presidential elections.[32] After she left the company, she disclosed private documents from Meta showing that the company was aware of the harmful effects its services were having on society[33] – and claiming that they consciously disregard those findings for profit's sake.[34] Following the leaks, she was invited to testify before the US congress on October 5th, 2021. The second whistle blower we look at here is Peiter Zatko (male). Also known as "Mudge", Mr. Zatko was Twitter's security chief from 2020 to 2022 – and was hired directly by then-CEO Jack Dorsey. After the company terminated him, in late August 2022, he raised allegations against Twitter, regarding the latter's deficiency in handling

[32] Milmo, Dan, and Dan Milmo Global technology editor. 'How Losing a Friend to Misinformation Drove Facebook Whistleblower'. *The Guardian*, 4 October 2021, sec. Technology. https://www.theguardian.com/technology/2021/oct/04/how-friend-lost-to-misinformation-drove-facebook-whistleblower-frances-haugen. Accessed April 20th 2023.

[33] This resulted in a series of articles published mostly in the Wall Street Journal. For a summary see Milmo, Dan, and Dan Milmo Global technology editor. 'Facebook "Tearing Our Societies Apart": Key Excerpts from a Whistleblower'. *The Guardian*, 4 October 2021, sec. Technology. https://www.theguardian.com/technology/2021/oct/04/facebook-tearing-our-societies-apart-key-excerpts-from-a-whistleblower-frances-haugen. Accessed April 20th 2023. For the original series see *Wall Street Journal*. 'The Facebook Files'. 1 October 2021, sec. Tech. https://www.wsj.com/articles/the-facebook-files-11631713039. Accessed April 20th 2023.

[34] Paul, Kari, and Dan Milmo. 'Facebook Putting Profit before Public Good, Says Whistleblower Frances Haugen'. *The Guardian*, 4 October 2021, sec. Technology. https://www.theguardian.com/technology/2021/oct/03/former-facebook-employee-frances-haugen-identifies-herself-as-whistleblower. Accessed April 20th 2023.

users' information and violation of US legislation.[35] His allegations were crucial in the leadup to Elon Musk's acquisition of Twitter later that same year. Zatko was invited to testify before the US congress on September 13th, 2022.

*3.2 Data Collection* - The investigation was conducted on Twitter, the online social media usually employed as a proxy for the online public sphere. The data was collected using the scrape function of the Minet Python library, a scraping tool developed by the Sciences Po Médialab,[36] using the text strings "Frances Haugen"; "Peiter Zatko"; and "#mudge" – the latter employed to account for Mr. Zatko's nickname which has been extensively used in the online debate around his leaks. The timeframe for the data collection was of around two months after the news broke for both cases – more precisely between October 1st and November 30th, 2021, in the case of Haugen, and from August 15th to October 15th, 2022, for Zatko. This search yielded a total of 23 334 tweets for Haugen and 2 896 for Zatko, reflecting the different sizes of the two scandals.

**Table 1. Summary of Data Collection**

| Whistle-Blower | Search String | Timeframe | Number of Tweets |
|---|---|---|---|
| Frances Haugen | "Frances Haugen" | 01/10/2021 – 30/11/2021 | 23 334 |
| Peiter Zatko | "Peiter Zatko"; "#mudge" | 15/08/2022 – 15/10/2022 | 2 896 |

*3.3 Methodology*[37] - In order to investigate how gender influences the public debate in our case, this study looks at different aspects of the Twitter discourse. First, we look at the sentiment of the debate. The interest here is to detect whether one of the two cases had an overall more negative sentiment than the other – often related to more abusive language. To do so, we employ sentiment analysis through the use of the *SpaCy*

[35] Contrarily to Haugen, he filed a formal whistle-blower complaint to the US congress which were later published by the press. For a summary of the allegations see Vincent, James. 'Twitter's Former Security Chief Says Company Lied about Bots and Safety'. *The Verge*, 23 August 2022. https://www.theverge.com/2022/8/23/23317857/twitter-whistleblower-zatko-security-spam-safety.

[36] For documentation see https://github.com/medialab/minet/blob/master/docs/twitter.md. Accessed April 20th 2023.

[37] For access to the code used in this research see https://github.com/giomagg/Digital_Inequalities_SPRING_2023.

python library.[38] This is a pipeline of a trained natural language processing tool which yields results ranging between -1 (negative sentiment) and 1 (positive sentiment).

Secondly, we turn to the identification of sexism and gender-based harassment in the tweets contributing to the online discourse. To identify sexism this study employs a supervised-learning algorithm trained specifically for this purpose. I employed the *AugmentedSocialScientist* library[39] to condition the pre-trained BERT language model on the labelled dataset developed by Mattia Samory (2021).[40] The algorithm yielded satisfying results, identifying sexism in tweets with an accuracy of 93%.[41] In a less virtuous fashion, I employed word count to identify gender-based harassment using the dictionary developed by Rezvan et al. in 2018.[42]

Finally, the study turns to the analysis of the topics around which the Twitter debate in the two cases developed. To study this, I employed the *BERTopic* topic modelling library.[43] The way this works is by first conditioning the BERT model's embeddings on the corpus of documents of interest (in this case the scraped tweets). Secondly, employing UMAP and HDBSAN the dimensionality of the embeddings is reduced and then clustered based on semantic similarity. Third, topics are created out of the semantic clusters. In the end, this method allows to extract and analyse the main topics present in a given corpus.

[38] For documentation see https://spacy.io/universe/project/spacy-textblob.

[39] For documentation see https://github.com/rubingshen/AugmentedSocialScientist.

[40] Samory, Mattia. 'The "Call Me Sexist but" Dataset (CMSB)'. GESIS Data Archive, 2021. https://doi.org/10.7802/2251.

[41] More specifically, the algorithm was better at detecting cases of non-sexism (does so with 96% accuracy), while it performs worse in detecting sexism (74% accuracy). This is due to the fact that the training dataset contains more cases of non-sexism than cases of sexism, thus allowing the algorithm to detect the former better than the latter. However, overall, the algorithm performs quite well with an accuracy of 93%.

[42] Rezvan, Mohammadreza, Saeedeh Shekarpour, Lakshika Balasuriya, Krishnaprasad Thirunarayan, Valerie L. Shalin, and Amit Sheth. 'A Quality Type-Aware Annotated Corpus and Lexicon for Harassment Research'. In *Proceedings of the 10th ACM Conference on Web Science*, 33–36. WebSci '18. New York, NY, USA: Association for Computing Machinery, 2018. https://doi.org/10.1145/3201064.3201103.

[43] For documentation see https://maartengr.github.io/BERTopic/index.html.

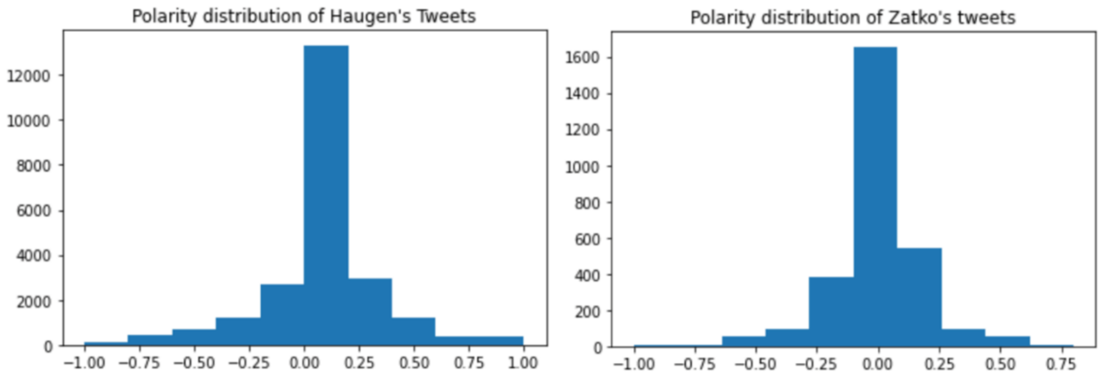

**Figure 2 – Comparison of Zatko's and Haugen's Polarity distribution**

**4. Findings-** *4.1. Sentiment Analysis* - In the two cases, the sentiment of the discourse was on average emotionally neutral – meaning that there was a similar number of tweets displaying negative and positive sentiment. In fact, the average polarity for Haugen was 0.043 while that for Zatko was 0.0026 – where polarity ranges between -1 and 1. Moreover, the distribution and variation of polarity over time (see figure 1-2) is comparable in the two cases, with no clear trend indicating substantial differences between the two. As we can see, both distributions follow a somewhat normal one centred around their mean. The only difference between the two is found in their standard deviation, which in the case of Haugen is slightly higher (0.25) than in that of Zatko (0.17).

*4.2. Identification of harassment and sexism* - Turning now to the detection of gender-based harassment, the analysis – based on Rezvan et a. (2018) – finds some differences between the two cases. In fact, while in the case of Peiter Zatko the percentage of tweets containing harassment is 6.5%, for Frances Haugen it is in double digits: 10.2%. Figure 3 shows a breakdown of the most common words employed in this regard. Moreover, no substantial difference was detected regarding sexism. The algorithm

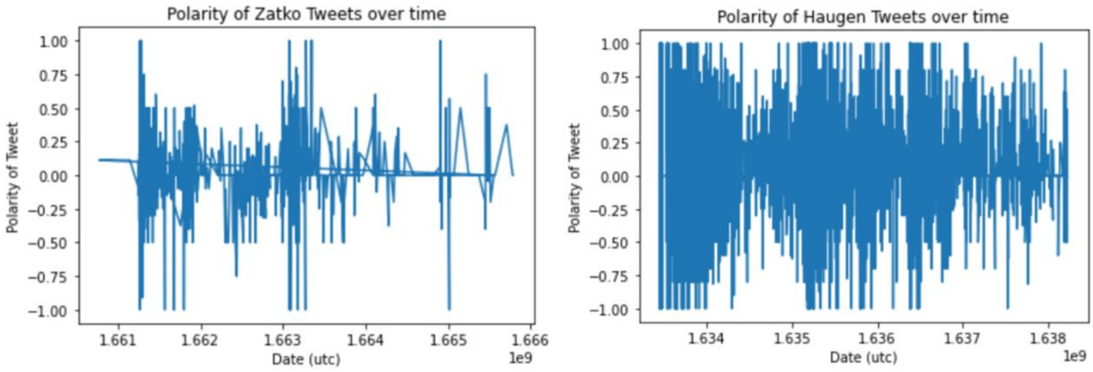

**Figure 1 – Comparison of Zatko's and Haugen's polarity variation over time**

detected 19 cases of sexism in Haugen's datasets and 0 in Zatko' – respectively 0.00082% and 0% of all tweets. Table 2 here below summarises these findings.

**Table 2. Summary of Sexism and Harassment Statistics**

| **Whistle-blower** | **Haugen** | **Zatko** |
|---|---|---|
| **Sexism** | 19 tweets (0.00082%) | 0 tweets |
| **Harassment** | 10.2% | 6.5% |

*4.3. Topic Modelling* - Finally, we turn to look at the outputs of the topic modelling. Overall, the gender of the two whistle-blowers was not a prevalent topic of discussion – i.e., it was not detected in the most prevalent topics in neither one nor the other dataset (see figure 4 and 5).

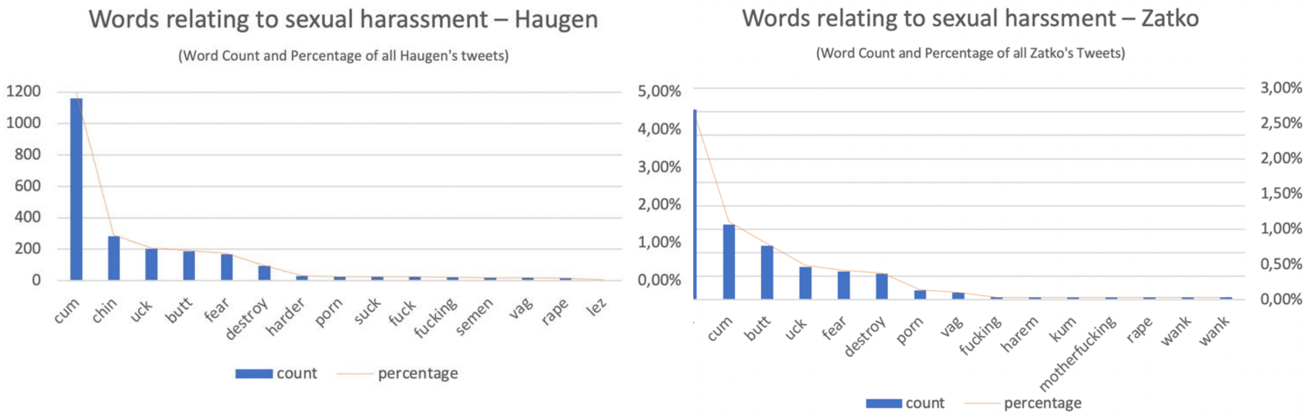

**Figure 3 – Comparison of Zatko's and Haugen's words related to sexual harassment** (Please note: "uck" is internet slang to indicate oral sex)

Gendered words relating to the identity of the whistle-blower appear only in one topic cluster (number 13, figure 4). The discourse around the gender of Haugen is related to positive adjectives within the cluster. The topics around which the debate was conducted however mostly dealt with political topic related to the leaked documents as well as around the congressional hearings and testimonies of the two. In fact, in the case of Haugen we find many topics dealing with the mental health effects of the use of social media, as well as the global impact of social media on developing countries. For Zatko, the focus of the conversation was more on its testimony, Elon Musk's takeover of Twitter – for which Zatko's revelations were highly relevant –, and his revelations.

**5. Discussion, Limitations, and Conclusion -** Although gender-based harassment is still present in the online environment and displays different magnitudes for male and female, this analysis has shown that for whistle-blower the discourse tends to focus prevalently on the content of their leaks rather than on the gender aspect. Moreover, no substantial difference was detected concerning the polarity of the two discussions, both of which are overall neutral. The absence of gender elements in the discourse about whistle-blowers' leaks suggests that the these aspects would not play into their perceived support. In turn, this points to the fact that we will not see substantial differences in the actors bringing to the public sphere information which are in the public interest. Or at least, that disincentives do not come from the online discussion.

This study has some limitations. First, the supervised sexism-detection algorithm could be better trained on a wider dataset to increase its accuracy – although no substantial difference with the results here presented is expected. Secondly, the method gender-based harassment could be improved, maybe also through the use of supervised leaning. Thirdly, the claims presented in this paper are specific to the comparison under consideration. There is the need to test these results on a higher number whistle-blower cases, also by looking at how the debate plays out on different social media platforms as well as in the press. This would also allow to test different kinds of identities that go beyond the binary distinction between male and female. In this regard, future research might take into account whistle-blowers from the LGBTQ+ such as Chelsea Manning. Finally, future research should also take into account the community specific nature of online conversation – not taken into account here due to space constraints. What this means is conducting social network analysis to identify different communities and conduct topic modelling – in a similar fashion to how it has been implemented here – on a community basis. Employing social network analysis would also allow for the identification of prevalent actors pushing certain topics and promoting specific narrative, enabling for a more in debt analysis of the identities of these actors.

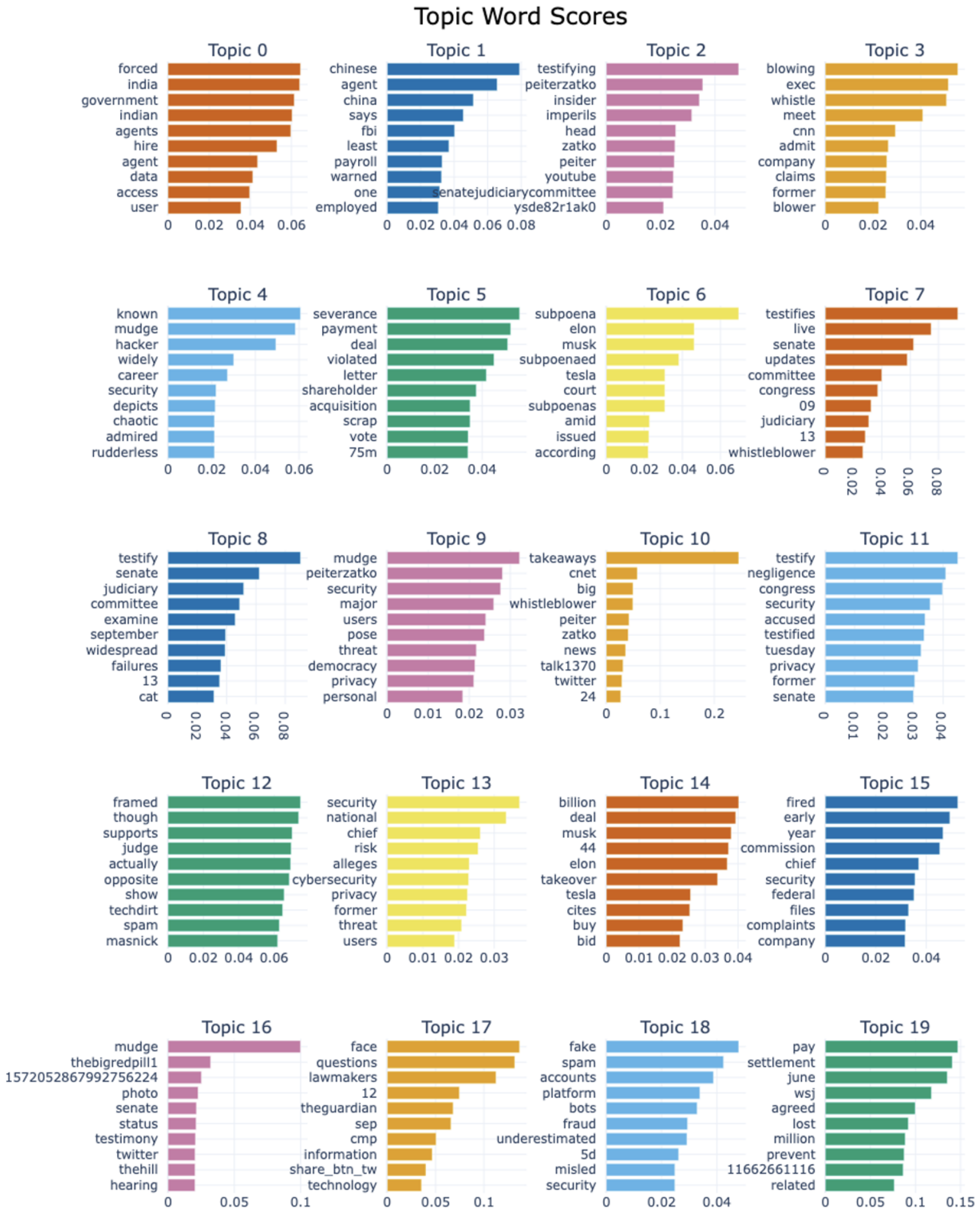

**Figure 5 – Topic Model for Peiter Zatko (BERTopic)**