Can there be a feminist text analysis?

Reflections on science, digital humanities, and text analysis methods and applications

Livia Clarete
Professor Lisa Rhody
Feminist Text Analysis (DHUM72500)
26 May 2023
Github repo here

Introduction

"Data is the new oil."[1]. This famous quote was first attributed to Clive Humby in 2006, the British mathematician and data scientist who openly discussed the importance that data plays in the modern economy. The amount of information produced in 2022 is estimated at 64.2 zettabytes, sixty times greater than the 1.2 zettabytes from 2010. By 2025, society is expected to produce 180 zettabytes[2]. Not only is the amount of information growing, but even the metrics to measure it: a zettabyte represents 1 trillion gigabytes. Most of the new data is unstructured. It comes as texts, images, audios, and other sources of rich data. The question now is how to extract meaningful information and learn from this material. Recently, this idea from early 2000 has been challenging: data is not the new oil. Time is[3]. This is our most valuable currency. Humans are *still* better in judgment, but our time and attention are limited to manually processing them (Nguyen et al, 2019).

Data analytics methods have experienced a significant expansion beyond their traditional application in the business and corporate sectors. This expansion has permeated academia, resulting in the emergence of new fields of interdisciplinary studies, such as digital humanities. This approach has been leveraging data-driven insights and boosting new methodologies in humanities studies. Computational methods have been established as an important set of tools to analyze social and cultural patterns in the digital world, from exploring hate speech in social media to authorship of The Federalist Papers (Nguyen et al, 2019).

Humanities researchers are at the forefront of pushing the boundaries of social studies in various areas. Blevins et al. (2021) delve into the realm of social media communities. They explore new concepts, like the "digital street" by shedding light on how gang-involved youths in Chicago propagate a new dimension of violence through online platforms. Noble (2018) deeper into search engines, gender, and race. She questioned Google's neutrality by providing evidence supporting that search results for "black girls" are often sexualized and objectifying. Noble argues that this is not an isolated incident, but rather a symptom of a larger problem with the way that search engines are designed and operated. Beyond their contributions with social insights about the structure and conjuncture of today's society, the aforementioned authors bring new methodologies for analyzing text and digital footprints.

---

[1] Chazan, Guy. "'Data Is the New Oil . . . Who's Going to Own It?'" Financial Times, November 16, 2016.
[2] Statista. 2021. "Data Created Worldwide 2010-2025 | Statista." Statista. Statista. June 7, 2021.
[3] Forbes. (2021, July 15). Data Isn't The New Oil — Time Is.

When science overcomes neutrality

The notion of the importance of building reliable methods in Social Science dates back to the discipline's origins in late 19th-century Europe. The French philosopher Auguste Comte[4] argued that Sociology should adopt the positive method to establish itself as a way to be considered a science, like hard sciences such as Biology and Mathematics. Émile Durkheim[5], often regarded as the founder of Sociology as an academic discipline, emphasized the importance of neutrality in Social Science research. He advocated for researchers to avoid expressing their personal emotions and beliefs, focusing instead on presenting the facts. This approach was based on the belief that humans are rational beings. Recent studies have been proving it to be wrong. Kahneman[6] has written extensively about the limitations of human rationality: humans are biased by nature and we should be aware of that. The discussion in academia about the concept of neutrality has been changing. As the 500 Women Scientists Leadership mentioned, "*Ignoring science's legacy of racism or a wider culture shaped by white supremacy doesn't make scientists "objective*""[7]. On the contrary, as pointed out by Paulo Freire, being silenced in the face of inequality is not neutrality, but rather a way that reinforces unequal power structures[8].

"Man is by nature a Political animal" (Aristotle, 1998). People, both man and woman, are gregorian beings that inheritably are capable of Political associations with others and moral reasoning[9]. However, the question of whether science should be involved in Politics is still under development in our society. Nature, one of the most prestigious scientific journals, launched a series of podcasts discussing why a journal of science covers politics - mostly answering comments from social media users questioning their content taking part in the recent threats to the Democratic system worldwide and the decreasing trust in vaccines and scientific knowledge in general. They replied explaining that Nature has had close relations with politics since their beginning, when in 1869 they vindicated that science should be taught at schools. In the following year, they promoted scientific education for women. This editorial changes to a more apolitical approach during World War II, which limits their bandwidth to discuss controversial topics. Since then, they were taking mild approaches to discussing politics, until the moment in which the scientific realm was put in danger with the increasing social discredit[10].

Common sense may think that taking part and overcoming neutrality in research leads to individual judgment. Science, especially social science, can be political, but it is not condemnatory. Taking an example of the Peer to Peer (P2P) economy, Doleac and Stein (2013) found that when the same item (an iPod Nano) was held by a dark-skinned hand, compared to a light-skinned hand, advertisements resulted in fewer responses, fewer and lower offers. Ayres et al (2015) performed an experiment that showed cards held by either a dark-skinned hand or a light-skinned hand. Cards held by African-American hands sold for approximately 20% less compared to Caucasian ones. Edelman and Luca (2017) concluded that travelers with "distinctly African-American names" were 16% more likely to be rejected by Airbnb hosts than "identical guests with distinctly white names". Zhang et al. (2021) showed that black hosts earn less on the platform compared to their white counterparts. Research exploring racism, sexism, and other prejudices is focused on uncovering patterns

[4] Comte - Positivistic Approach, 2023
[5] The Rules of Sociological Method (1895)
[6] Kahneman, Daniel. Thinking, Fast and Slow.
[7] Leadership, 500 Women Scientists. n.d. "Silence Is Never Neutral; Neither Is Science." Scientific American Blog Network
[8] Freire, Paulo. Pedagogy of the Oppressed.
[9] "Higher" and "Lower" Political Animals: A Critical Analysis of Aristotle's Account of the Political Animal,
[10] 'Stick to the science': when science gets political

in society, rather than theorizing about any sort of individual character deviation. Instead, these results foster media and civil society organizations to push for change and policy reviews. In the Airbnb case, New York Times journalist, Katie Benner, backed her article with scientific results to pressure Airbnb to adopt rules to fight against discrimination[11], including a quote from Brian Chesky, Airbnb's chief executive stating that the company needed to do better in dealing with this issue.

A feminist text analysis

In the humanities field, the feminist approach is prominent in discussions on how to build an equal society. Sara Ahmed (2017) defends a feminist perspective within the scientific realm of social science. Differently from what she calls 'moral feminism,' a feminist perspective engages in asking ethical questions about how to live better in an unjust and unequal world, rather than focusing solely on individual moralism. According to Ahmed, the most challenging questions revolve around explaining violence, inequality, and injustice. These include inquiries about what it means to be a feminist, how to address and change gender stereotypes, how to create a more equitable society, and how to foster social (institutional) transformation. Ahmed goes further by reflecting on the ways gender is constructed and represented in text, and the impact this has on society's understanding of gender stereotypes.

Eckert and McConnell-Ginet (2003) propose a comprehensive framework for gender text analysis that emphasizes the role of language in constructing and performing gender identities. They argue that gender is not fixed, but rather a negotiated and performed aspect influenced by language use. Their three-part framework includes genderlects, which encompass linguistic features associated with specific genders; gender identities, which represent social categories such as femininity, masculinity, and transgenderism; and gender performance, which explores how individuals employ language to construct and express their gender identities. This analysis aims to challenge traditional gender norms and promote inclusivity in society. Gender text analysis offers various practical applications. Firstly, it can help identify instances where language is used to stereotype and marginalize women and other marginalized groups. By understanding these patterns, interventions can be developed to combat discrimination and promote equality. Secondly, gender text analysis enables the examination of how language can be harnessed to advance gender equality and social justice, allowing for the creation of more inclusive communication practices. Lastly, this approach can inform the development of educational materials to enhance people's awareness of how language influences the construction and performance of gender identities.

The computational method of text analysis - a feminist approach

Building a method to ensure quality and reliable results is crucial. It includes exploring a conceptual framework, posing research questions, collecting, compiling, and validating data sources, modeling and analyzing the data, operationalizing the method, and interpreting the results. Open-source and free tools, such as Python and Colab, make this process easier, more automatic, and reproducible. In this section, we will be exploring the learnings from the coursework in text processing using NLTK.

[11] Airbnb Adopts Rules to Fight Discrimination by Its Hosts

Week 2: We were introduced to the concept of the notebook, either through Jupyter or Google Colab. Additionally, we explored the NLTK Python Library and worked on tasks such as tokenizing a sentence and performing POS tagging on the tokens. Two functions were used: nltk.word_tokenize(sentence), and nltk.pos_tag(tokens).

Week 3: We imported some books from the NLTK corpus library, then used the "concordance", "common_contexts", and "similar" functions to examine particular words to explore their usage patterns, shared contexts, and linguistic similarities. We also counted the number of words in the corpus, checked categories, and explored some visualizations, such as plotting frequencies in a line chart and a dispersion plot (dispersion_plot function). We discussed the concept of "material negotiations" according to McGann. It encompasses the practical and tangible aspects of cultural production that facilitate symbolic exchange. This implies that symbolic exchanges are intricately linked to their material contexts, starting from the act of reading itself. It involves considering the cultural elements that individuals possess, allowing them to comprehend and interpret the text. In the case of concordance analysis, the symbolic exchange commences with the utilization of a computer and an algorithm to locate instances of the word "monstrous" within the text of Moby Dick. The fundamental basis of this process lies in the material resources of computers and the supporting technology. Furthermore, the cultural background of the researchers who will analyze the results plays a significant role. This same process applies to similarity analysis as well.

Week 4: We were introduced to more advanced Python functions, including for loops, if/else statements, and list comprehensions, which allow for more efficient code creation. In addition, we conducted an analysis in week 3 using dispensary plots and standard text, focusing on the novel Sense and Sensibility. The dispersion plot revealed the presence of four characters, with Willoughby and Edward consistently appearing throughout the narrative, indicating their roles as lead characters. Elinor, as the most mentioned character, held significant importance. Marianne and Elinor appeared at intervals, suggesting secondary roles, and their appearances did not directly coincide, indicating an unrelated relationship. However, Elinor and Edward frequently coincided in their appearances, indicating a possible connection as a couple. Similarly, Marianne and Willoughby's appearances coincided, suggesting a relationship between the two characters. Lastly, we explored new research questions that could be answered using text analysis, such as identifying the tone, context, and style of the text.

Week 5: We delved deep into our research questions and conducted an analysis using a series of books called Blueprints. To begin, we imported the CSV file using the 'read_csv' function from the Pandas library. Next, we created various features based on the text length, addressed any missing values, performed calculations, and visualized the results using charts. The visualization included plots showcasing the usage of n-grams over time, frequency charts, and word clouds. These visual representations provided valuable insights into the data and allowed us to explore patterns and trends.

Week 7: We discussed the importance of data collection and the criteria for determining 'good data.' One automated method for extracting data is by utilizing Python libraries such as Requests and BeautifulSoup. These libraries enable the downloading of web pages and the parsing and extraction of data from HTML and XML files. They allow for the removal of specific elements like tables, paragraphs, or headings from HTML

files, facilitating further analysis or visualization. In addition, we explored Rob Kitchin's (2014) criteria for evaluating 'good data,' which include accuracy, reliability, validity, timeliness, and accessibility. These criteria are essential for ensuring the quality of the data and its suitability for analysis purposes.

Week 8: Discussion about the concept and applications of APIs using Requests. Roughly, an API is a software that connects with another software and pulls data from it. The data was collected from chroniclingamerica. The request was made and retrieved a JSON data format. Then, it was transposed into a pandas data frame.

Week 10: This notebook was not an assignment, but an explanatory notebook covering various concepts such as vectors, features, and similarity; feature engineering and syntactic similarity; data preparation; a bag of words; count vectorizer; TF-IDF (term frequency-inverse document frequency) method, which helps identify important words in a collection of documents; dimension reduction; calculating cosine similarities; n-grams; lemmas; and finding similar and related words. The second notebook of week 10 applied these concepts to the IMDB dataset using TensorFlow. The data was imported, and the model was trained using Keras, a TensorFlow API. The model was built in a sequential manner, where each layer is added one after another in a linear fashion. The first layer was a hub layer, which was a pre-trained embedding layer from TensorFlow Hub. The second layer was a dense layer with 16 neurons and a ReLU activation function. Finally, there was a final denser layer with one neuron. The model's performance was evaluated based on accuracy, and the results were plotted on a chart.

Week 11: This was not an assignment, but rather a notebook where supervised classification was performed using sklearn models, such as SVC and LinearSVC. The data was imported using Pandas, and exploratory data analysis was conducted using bar charts. NaN values were dropped, and a train-test split was performed before applying TF-IDF. The models were trained, and the y variable was predicted, followed by evaluation using accuracy score, precision, and recall. Cross-validation was performed using the cross_val_score function, and the hyperparameters were tuned using Grid Search.

From a feminist perspective, the reflection involved the data collection process, which plays a crucial role in the overall data process and analysis pipeline. High-quality data is essential for conducting quality analysis and should be collected in an unbiased, diverse, and representative manner to ensure integrity and reliability. Addressing potential biases and social inequalities within the Java workplace requires evaluating the model's impact and considering any issues that may arise from insufficient training. Ethical considerations come into play when assessing the implications of using classification models and weighing them against potential harms, with a focus on enhancing user and employee experiences. To enhance efficiency, the selection of the most effective model is paramount, and while energy consumption may be challenging to measure accurately, time processing can serve as an evaluation metric. By conducting an initial analysis using a sample dataset, practitioners can assess the cost-benefit ratio and subsequently employ the optimized model for a full analysis. Additionally, creating Key Performance Indicators (KPI) efficiency metrics can aid both practitioners and academics in utilizing language models more effectively.

Week 14: Topic analysis with Gensim applied to state-of-the-union CSV data. The process involved creating a dictionary and filtering out extreme words, then creating a corpus using the doc2bow function. Then, a TF-IDF function was applied to the data before building two models: LSI and LDA models. The topics were printed, then visualized using both bar charts and pyLDAvis - specifically for visualizing topics.

Conclusion

It is evident that a feminist text analysis holds significant value and relevance in today's society. The feminist perspective in social sciences emphasizes ethical questions surrounding inequality, injustice, and gender stereotypes. This perspective recognizes the role of language and representation in constructing gender identities and shaping societal understanding. Gender text analysis provides a comprehensive framework to analyze language's influence on gender, challenging traditional norms and promoting inclusivity. Such analysis can help identify and shed light on discrimination and enhance awareness of language's role in constructing gender identities. By delving deep into research questions and using advanced analysis techniques, researchers can explore various aspects of text analysis and its implications.

Bibliography

Nguyen, Dong, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, and Jane Winters. "How we do things with words: Analyzing text as social and cultural data" arXiv:1907.01468v1 [cs.CL] 2 Jul 2019. Retrieved from https://arxiv.org/pdf/1907.01468v1.pdf

Safiya Umoja Noble, "Searching for Black Girls" Algorithms of Oppression: How Search Engines Reinforce Racism. New York: NYU Press, 2018 - View a talk based on this book: https://youtu.be/iRVZozEEWlE

Kahneman, Daniel. Thinking, Fast and Slow. New York: Farrar, Straus and Giroux, 2011.

Freire, Paulo. Pedagogy of the Oppressed. London: Penguin Classics, 2017.

Aristotle and T. A. 1899-1961. Sinclair. 1962. The Politics. Baltimore, Penguin Books.

Eckert, Penelope, and Sally McConnell-Ginet. 2013. Language and Gender. 2nd Edition. Cambridge UP: Cambridge. pp 1-36

Ahmed, Sara. "Introduction: Bringing Feminist Theory Home." Living a Feminist Life. Duke University Press, 2016. https://muse.jhu.edu/book/69122

Doleac, Jennifer L., and Luke C.D. Stein. 2013. "The Visible Hand: Race and Online Market Outcomes." The Economic Journal 123 (572): F469–92. https://doi.org/10.1111/ecoj.12082

Ayres, Ian, Mahzarin Banaji, and Christine Jolls. 2015. "Race Effects on EBay." The RAND Journal of Economics 46 (4): 891–917. https://doi.org/10.1111/1756-2171.12115

"The Rules of Sociological Method (1895)." n.d. Durkheim.uchicago.edu. https://durkheim.uchicago.edu/Summaries/rules.html

Edelman, Benjamin, Michael Luca, and Dan Svirsky. 2017. "Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment." American Economic Journal: Applied Economics 9 (2): 1–22. https://doi.org/10.1257/app.20160213.

Leadership, 500 Women Scientists. n.d. “Silence Is Never Neutral; Neither Is Science.” Scientific American Blog Network. https://blogs.scientificamerican.com/voices/silence-is-never-neutral-neither-is-science/.

Abbate, Cheryl E. 2016. “‘Higher’ and ‘Lower’ Political Animals: A Critical Analysis of Aristotle’s Account of the Political Animal.” Journal of Animal Ethics 6 (1): 54. https://doi.org/10.5406/janimalethics.6.1.0054

Kitchin, Rob. 2014. "Conceptualizing Data." _The Data Revolution_. NY: Sage Publishing,

“Comte - Positivistic Approach.” 2023. Pdx.edu. 2023. https://web.pdx.edu/~tothm/theory/DeadSoc/Comte/Comte%20-%20Positivistic%20Approach.htm

Forbes. (2021, July 15). Data Isn't The New Oil — Time Is. Retrieved from https://www.statista.com/statistics/871513/worldwide-data-created/

Financial Times. (2016, November 15). ‘Data is the new oil . . . who’s going to own it?’. Retrieved from https://www.ft.com/content/e548deac-856a-11e6-8897-2359a58ac7a5

Statista. (2022, September 8). Amount of data created, consumed, and stored 2010-2020, with forecasts to 2025. Retrieved from https://www.statista.com/statistics/871513/worldwide-data-created/

Chazan, Guy. 2016. “‘Data Is the New Oil . . . Who’s Going to Own It?’” Www.ft.com. November 16, 2016. https://www.ft.com/content/e548deac-856a-11e6-8897-2359a58ac7a5

Sinykin, Dan, and Edwin Roland. “Against Conglomeration: Nonprofit Publishing and American Literature After 1980.” Post45: Peer Reviewed, Apr. 2021. post45.org, https://post45.org/2021/04/against-conglomeration-nonprofit-publishing-and-american-literature-after-1980/

Howe, Nick. 2020. “‘Stick to the Science’: When Science Gets Political.” Nature, November. https://doi.org/10.1038/d41586-020-03067-w

Weekly assignments

Weekly topic	**Notebook**
Week 1 – Introductions	-
Week 2 – Sex, Gender, & Feminism	Clarete_Week2.ipynb
Week 3 – Text	Clarete_Week3.ipynb
Week 4 – Analysis	Clarete_Week4.ipynb
Week 5 – Research Questions	Clarete_Week5.ipynb
Week 6 – Data	-
Week 7 – Data	Clarete_Week7.ipynb
Week 8 – Conceptualization	Clarete_Week8.ipynb
Week 9 – Conceptualization	Clarete_Week10_theory.ipynb Clarete_week10_movie-review.ipynb
Week 10 – Operationalization	-

Week 11 – Operationalization	Clarete_Week11_classification.ipynb
Week 12 – Operationalization	-
Week 13 – Analysis	-
Week 14 – Analysis roundup	Clarete_Week14_topic_analysis.ipynb
Week 15 – Final	Portfolio - this document

Blog posts

- In Search of Zora/When Metadata Isn't Enough: Rescuing the Experiences of Black Women Through Statistical Modeling - Week 8: March 27 – Conceptualization - here
- Ulysses Interactive Visualization with scattertext (HDTMT) - here
- Book review: Ulysses by Numbers by Eric Bulson - here
- Weapons of Math Destruction - a deep dive into the recidivism algorithm - here
- Abstract: Unveiling the Patient Journey: A Gender Perspective on Chronic Disease-Centered Care - here

Contact information

Livia Clarete
livia.clarete@gmail.com