Word Association Thematic Analysis: Insight Discovery from the Social Web

Mike Thelwall
m.thelwall@wlv.ac.uk
Statistical Cybermetrics and Research Evaluation Group
University of Wolverhampton, UK

What is Word Association Thematic Analysis?

- A research method published in 2021
- Identifies themes in large collections of texts, such as tweets, YouTube comments, scientific abstracts.
- Addresses social science or humanities type research questions that can be answered by text analysis
- Combines automatic and manual stages
- Combines quantitative and qualitative methods
- Is inherently mixed methods
- Has supporting software Mozdeh (mozdeh.wlv.ac.uk)

What is Word Association Thematic Analysis (WATA)?

- A method to identify themes in sets of texts.
- Requirements:
  - A large set of texts: 10,000+.
  - Ideas for a comparison.
  - Free software Mozdeh to make comparisons. https://mozdeh.wlv.ac.uk.
- Output:
  - Themes related to the comparison.
  - Each theme is derived from a set of words identified from the comparison.

Example: International differences in Covid-19 vaccination tweets

- Taken from a paper, this is a WATA summary of themes in how UK Covid vaccine tweets differ from 7 other countries.
- UK used more positive terms (*fantastic*, *lovely*, *brilliant*, *pleased*, *well done*, *amazing*, & concluding kisses: *x* and *xx*).
  - US was *excited* (to get a vaccine) and Ireland was *delighted* (to get a vaccine).
- UK asked would vaccine *passports* be necessary for certain activities?
- UK asked would it be possible to *ease* the lockdown?
- Is the UK *Test and Trace* system an expensive failure?
- Vaccine hesitancy in the UK Black And Minority Ethnic (*BAME*) community.

WATA Requirement 1: Large set of texts

- Texts can be from any source, including:
  - Tweets (new Academic Research Twitter permissions)
  - YouTube comments
  - TripAdvisor reviews
  - Reddit posts
  - Titles, abstracts and keywords of academic papers
  - Reviews of academic papers
- Mozdeh can collect the tweets or YouTube comments or import others.
- 10,000 texts is the suggested minimum:
  - More texts give finer-grained themes, deeper insights.
  - More focused or longer texts also give finer-grained themes, deeper insights.

Requirement 2: A comparison

- The comparison may derive from the project goals.
  - **Gender**: Female v. male (v. Nonbinary if 100,000+ texts).
  - **Time**: Earlier v. later posts.
  - **Popularity**: Popular v. unpopular posts (retweets, likes, reward).
  - **Sentiment**: Positive v. negative posts.
  - **Country**: Country A v. Country B (or all other countries' posts).
  - **Topic**: Topic 1 v. topic 2 (e.g., lockdown v. "opening up").
- The comparison may be against a range of other topics.
  - E.g., ADHD tweets against 100 other medical conditions.

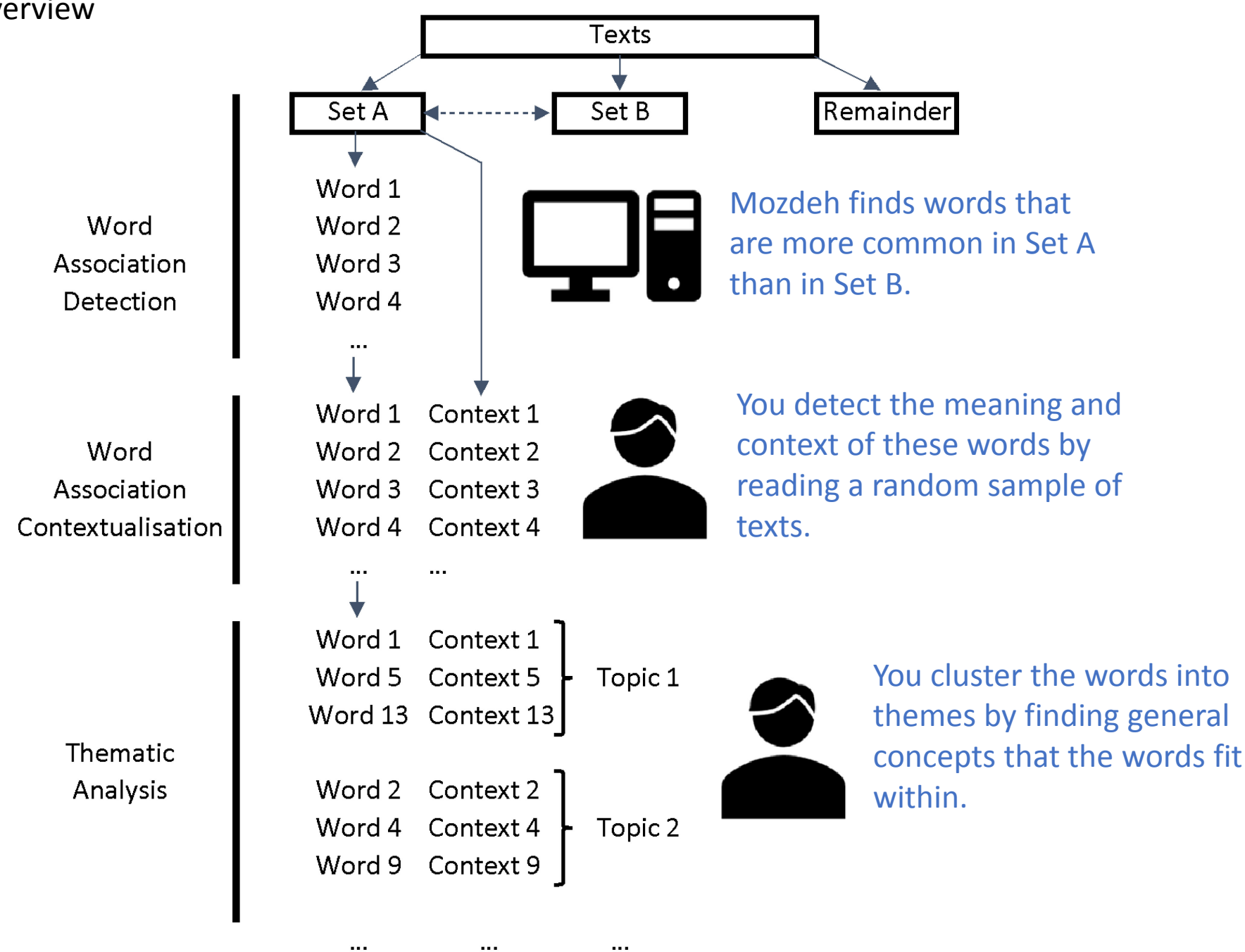
Examples of WATA studies

Study	Data	Comparison	Example finding
Gender differences in reactions to Covid-19	Tweets mentioning Covid-19	Female v. male	Females tweet more about safety, males more about politics.
Personal experiences of ADHD	Tweets about “my ADHD”	ADHD v. other disorders	The brain is discussed as a separate entity.
Evolution of #BlackLivesMatter during Covid-19	Covid-19 tweets about racism	Tweets in four different time periods.	The George Floyd killing led to tweets about systematic racism.
Self-presentation on Twitter	UK Twitter profiles	Female v. male v. nonbinary	Nonbinary profiles more likely to mention games and sexuality.
Discussions of bullying in YouTube	Comments on UK YouTube influencer videos	Bullying v. other topics	Strategies to support bullying victims include generalisation.

WATA Overview

1. Data collection: Mozdeh gathers/imports the texts to analyse.
2. Word Association Detection (WAD): Mozdeh identifies words that are more common each subset of the texts (e.g., male, female tweets) gathered in (1).
3. Word Association Contextualisation (WAC): You identify the typical meaning and context of each detected word by (2) by reading a random sample of texts containing it.
4. Thematic Analysis (TA): You identify themes by finding ways to group together the contextualised words from (3).

WATA Overview



1. Data collection

- Mozdeh can collect:
  - Tweets from keyword queries
  - Tweets from sets of users (e.g., all Maltese politicians' tweets)
  - YouTube comments for a set of videos or channels
- Mozdeh can import:
  - Plain text files
  - Tab-delimited plain text files
- See instructions at mozdeh.wlv.ac.uk for importing texts.

2. Word Association Detection

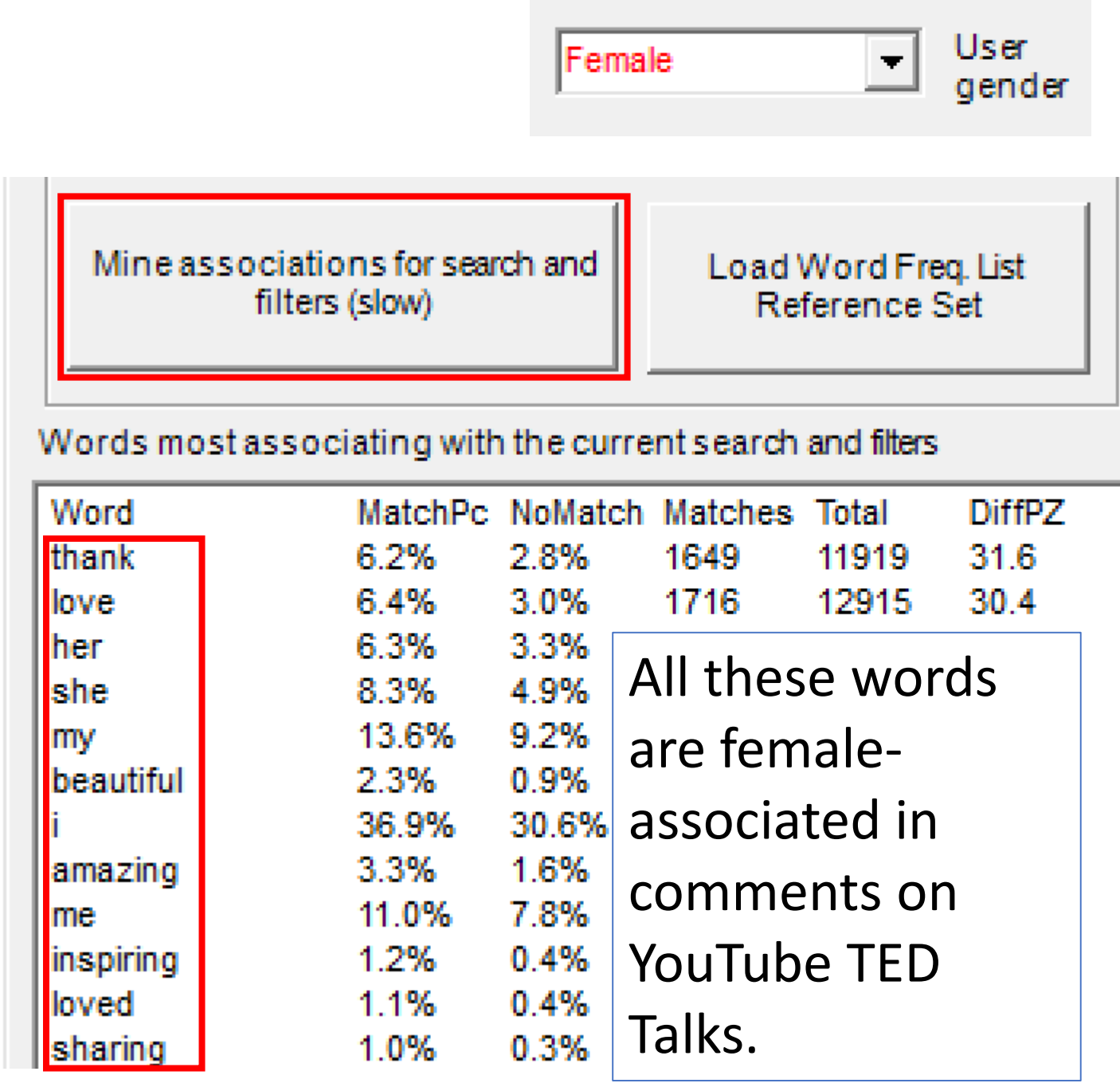
- Mozdeh's *Mine associations…* button detects words that occur more often in the specified set of texts than the remainder.
- A word **associates with a subset of texts** (e.g., female texts) if:
  a. The word occurs in **a higher percentage of** texts in the subset than the remainder, AND
  b. The difference between percentages is **statistically significant**.

WAD examples

- omg is female-associated in my data because it is (a) in 15% of female tweets but 10% of the rest and (b) Mozdeh reports the difference as statistically significant.
- wobble is not late-associated in my data because (a) it is in 4% of late comments but 5% of earlier comments.
- excellent is not popularity-associated in my data because whilst it is (a) in 10% of popular tweets and 9% of the rest, (b) Mozdeh reports the difference is not statistically significant.

WAD in Mozdeh

- Select the filters or enter the topic-defining query and click the *Mine associations…* button.
- Mozdeh will display a table of the top 1000 word associations.
- Terms with 1 to 3 stars are statistically significant.
- A chisquare test with Benjamini-Hochberg correction used to test statistical significance, protecting familywise error rates.



3. Word Association Contextualisation

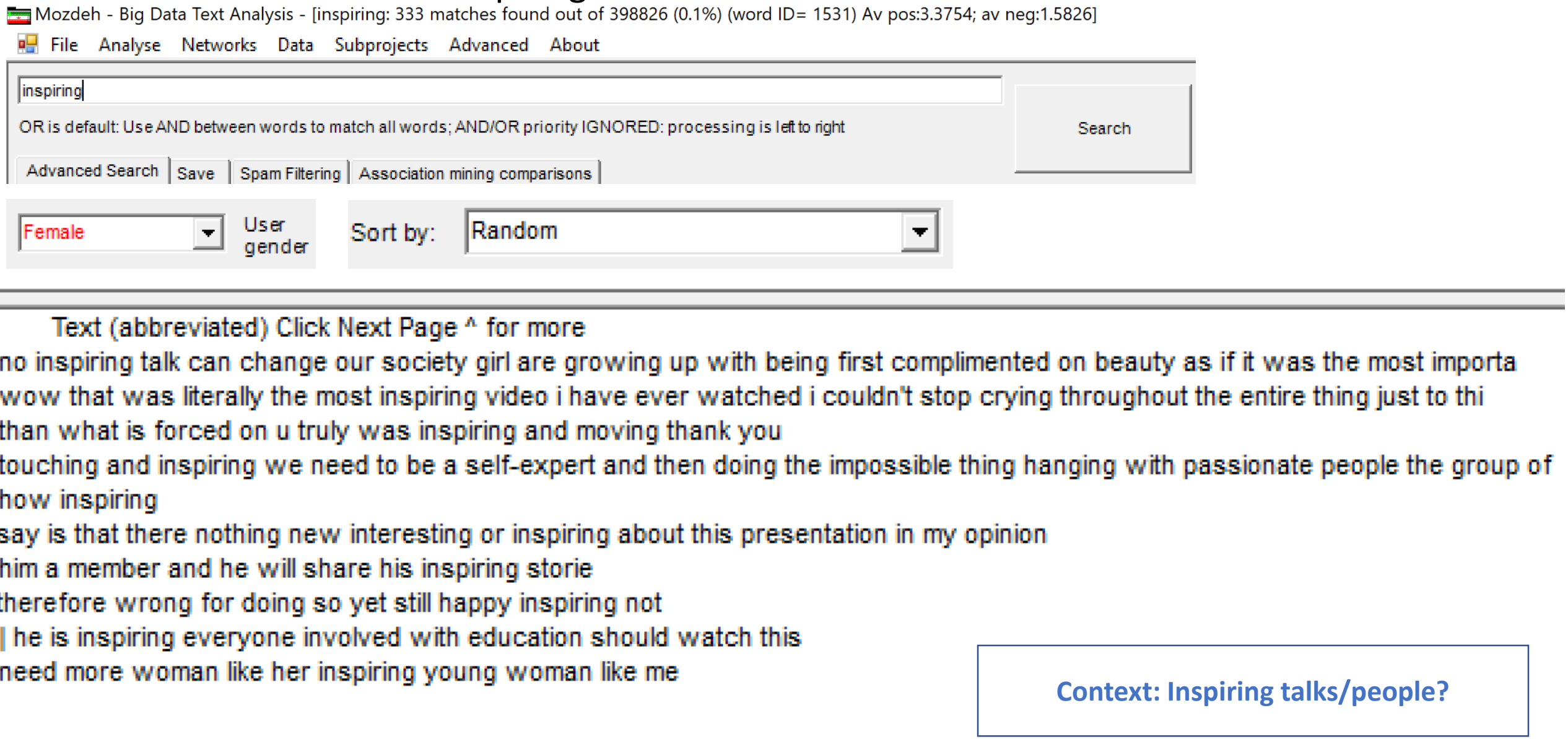
- WAD produces a list of associating terms but not their meaning or context.
- Detect the surface meaning of the term by reading 10+ texts containing it.
  - E.g., is “like” usually a comparator or a positive sentiment term?
- Detect the context of the term.
  - E.g., “congressional” usually used in the context of discussing congressional district voting results.
- Both meaning and context are needed to draw conclusions about a word association.
- A word’s context is the **most specific description that matches its most common use in the texts**.

KWIC WAC detect meaning and context

- Key Word In Context (KWIC) method (from corpus linguistics) is to read a random sample of 10+ texts containing a term to identify its meaning and context within these texts.
  - Read at least ten texts: more if important or the results unclear.
  - Identify the dominant meaning and contexts.
  - If there are no dominant meanings or contexts, discard the word.
- In Mozdeh, enter the term as a query, set corresponding filters and select the Random sort order to get a random selection of KWIC texts.

KWIC in Mozdeh example

Task: find the context of “inspiring” in female tweets.



Context: Inspiring talks/people?

KWIK in spreadsheets

From a set of tweets about Covid-19.

Word	Tweet
vulnerable	RT: Older relatives are vulnerable, we have to protect them.
vulnerable	I'm locking down with my daughter, who is on the vulnerable list.
vulnerable	My brother is vulnerable so I leave a weekly shop outside his door during co
vulnerable	Look after your vulnerable relatives so they don't get Covid-19, please!
coffee	I miss meeting friends for coffee during covid distancing.
coffee	I can shop online but can't chat with friends over coffee online. #Covid-19
coffee	I miss coffee meetings, restaurants, and pubs. #coronavirus
coffee	So sad not to catch up with you over coffee - see you after Covid-19!
coffee	The NCT coffee mornings were the best - can't believe they are all covid-can

Vulnerable context: People vulnerable to Covid-19?

Coffee context: Face-to-face coffee meetings?

KWIC WAC example

The context of love in the comments on TED talks below is:

Still love this one.

Absolutely love her.

I love her way to speak, to communicate, to say the right words. I just love her.

i love this man

This video is an essence of why I love Amanda.

I'd love to meet the speaker someday!

WAC Results

- After completing KWIC WAC for a specified number of word association terms (e.g., all significant terms, 100 terms), the result will be a list of meanings/contexts.
- In this example (racism on Twitter during Covid-19):
  - Two terms have unclear surface meanings, calling and slur
  - All terms have unclear contexts that needed KWIC to detect.

Term	**Meaning and context (Covid-19 tweets)**
disparities	Racial/ethnic disparities in coronavirus deaths.
anti-Asian	Opposition to anti-Asian racism due to Covid-19.
calling	Calling Covid-19 the "Chinese virus" is racist discussion.
slur	Arguing that "Chinese virus" is a "racial slur".

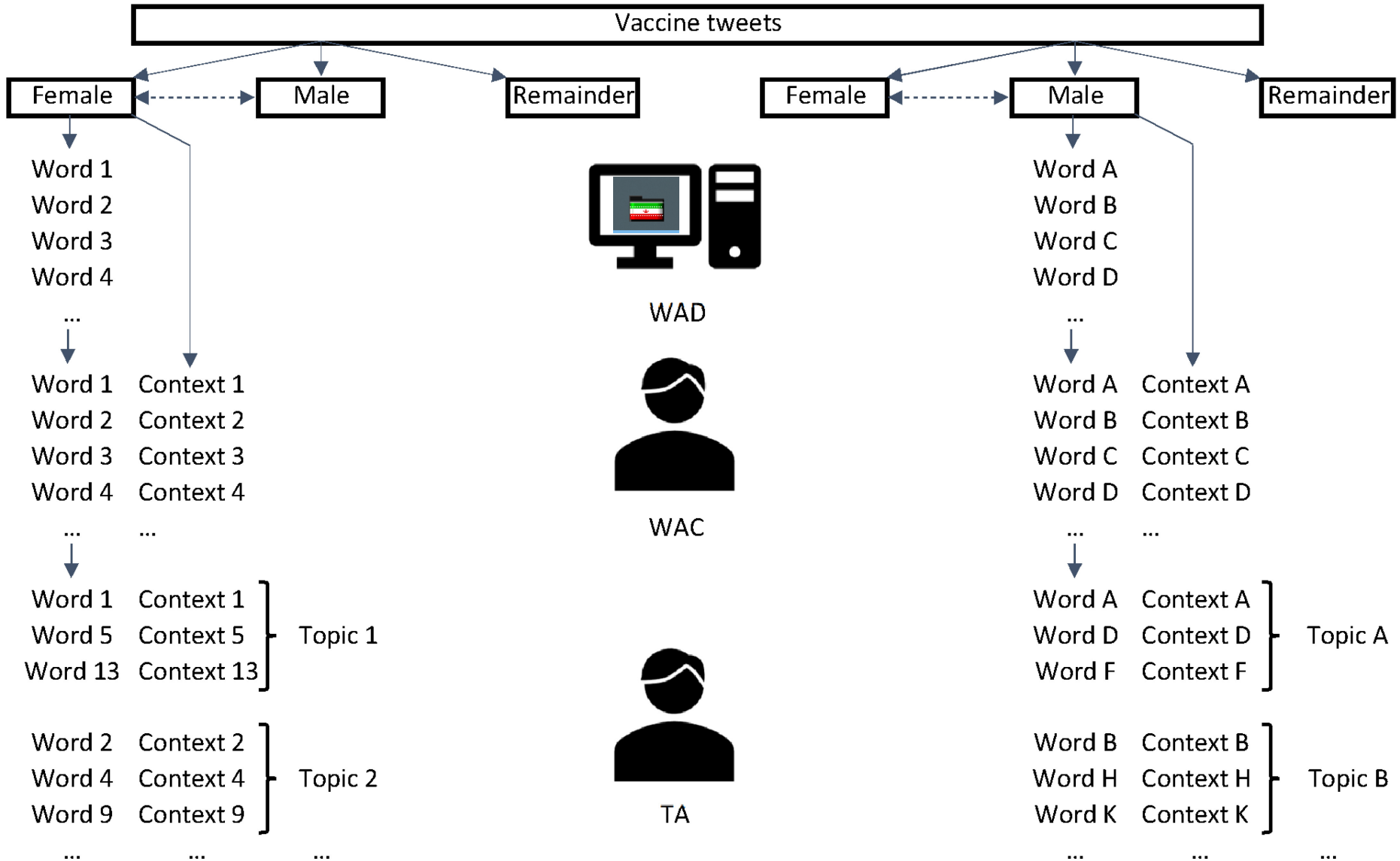
Word Association Analysis

- If there are few significant terms (e.g., <15) then thematic analysis is not be needed and the method is called *Word Association Analysis*: WAA=WAD+WAC.
- No need to cluster a few terms into themes.

Thematic Analysis (TA)

- The TA stage clusters the words into themes based on their contexts.
- A variant of the thematic analysis social science research method.
- Apply only after WAC and apply to the contexts.
- Tag each term with a *single* candidate theme generalizing its context.
  - Relevant to research goals, if possible.
- Try to find generalizing themes that can encompass multiple terms.
- Sort terms by theme (using a spreadsheet) and:
  - Look for opportunities to merge similar themes.
  - Check that all terms genuinely fit the theme.
- Recode/check themes on all terms when adjusting a theme.

Two Word Association Thematic Analyses for vaccine tweets to detect themes in tweets more posted by females than males (left), or more posted by males than females (right).



WATA Summary

- A method to identify themes in sets of texts.
- Needs 10,000+ texts, a comparison, and Mozdeh.
- Word Association Detection: Detect words associating with a topic or filter (e.g., gender, time, popularity, sentiment).
- Word Association Contextualisation: Identify the meaning and context of 100+ WAD words.
- Thematic Analysis: Cluster the terms into generalizing themes.

Two WATA studies

1. Bullying discussions in UK female influencers' YouTube comments.
2. "My ADHD hellbrain": A Twitter data science perspective on a behavioural disorder.

WATA Study 1: Bullying Discussions in UK Female Influencers' YouTube Comments

Thelwall, M. & Cash, S. (2021). Bullying discussions in UK female influencers' YouTube comments. *British Journal of Guidance and Counselling*, 49(3), 480-493.

Study 1: YouTube and Bullying

Motivation

- Female lifestyle influencers sometimes share personal issues as part of their friendly style.
- Bullying experiences and mental health issues are sometimes discussed.
- Comments on lifestyle influencers' videos may reveal attitudes towards bullying and styles of support.

YouTube GB

Search

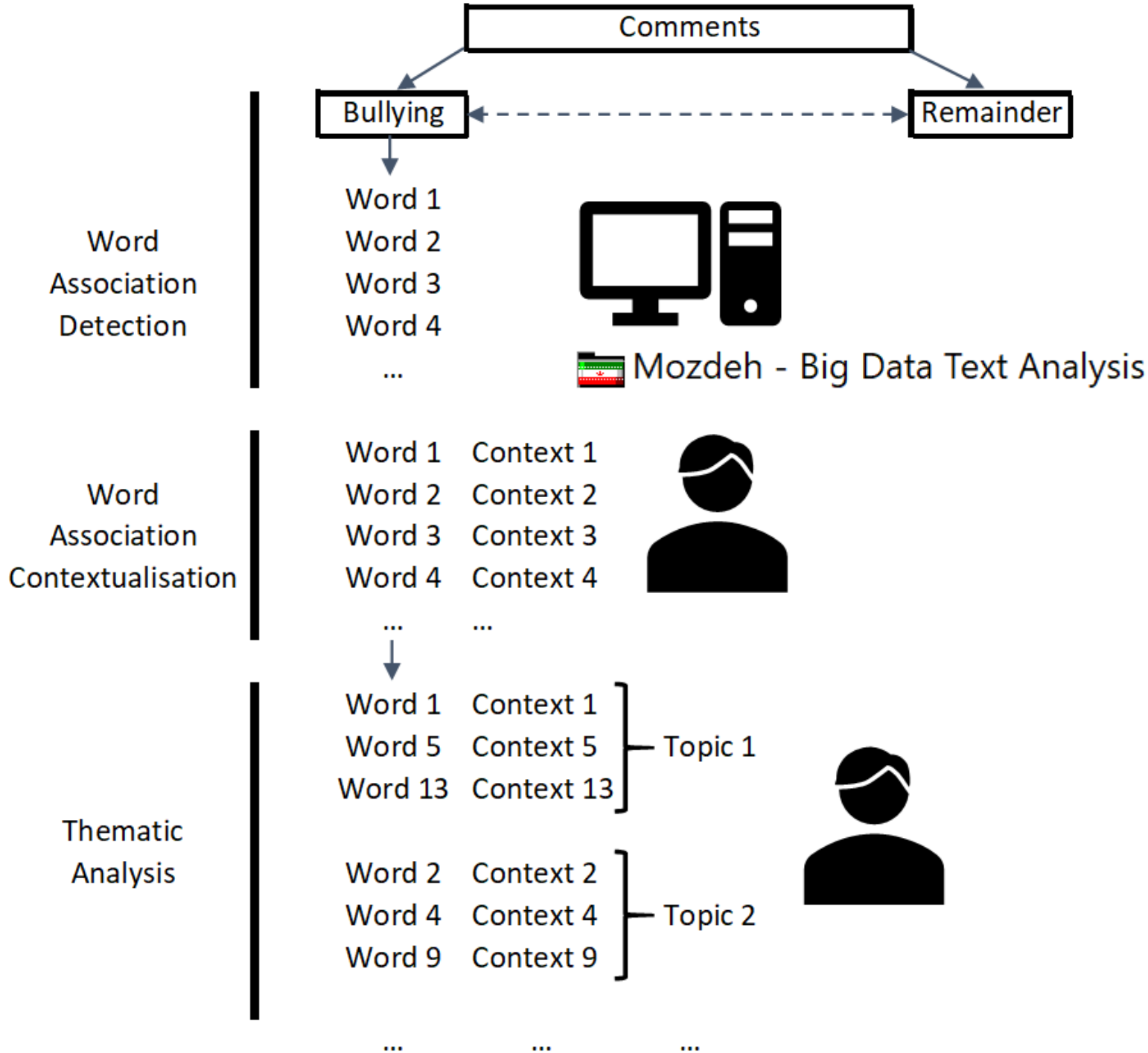
Get To Know Me | Q & A | Imogenation

86,035 views • 26 Feb 2017

3.3K 32 SHARE SAVE ...

Method: Word Association Thematic Analysis

- Objective: to get insights into bullying-related themes that are *different from* other discussions on influencer videos (to factor out the general tone of comments).
- Stage 1: Downloaded all comments on 34 popular UK female influencers' videos from YouTube with *Mozdeh*.
- Stage 2: Split the comments into two: bullying-related (*bullying bully bullied bullies cyberbully cyberbullied cyberbullies cyberbullying)* and the rest.
- Stage 3: Identified 100 words more used in the bullying set and organised them into themes.


Comments
Bullying
Remainder
Word 1
Word 2
Word 3
Word 4
...
Word Association Detection
Mozdeh - Big Data Text Analysis
Word 1 Context 1
Word 2 Context 2
Word 3 Context 3
Word 4 Context 4
... ...
Word Association Contextualisation
Word 1 Context 1
Word 5 Context 5
Word 13 Context 13
Topic 1
Word 2 Context 2
Word 4 Context 4
Word 9 Context 9
Topic 2
Thematic Analysis

Study 1 Results: Support for victims themes

- ***Thank you for sharing***: YouTubers and other commenters were **thanked** for telling about their experience in the comments section, or for opposing bullying.
- ***They are jealous***: There were discussions about **causes** of bullying, focusing on the bullies (*because, jealous, insecure*), or themselves as victims (*acne, [lack of] friends*) in personal stories.
- ***Terrible how people treat each other***: A common strategy was to **generalise or abstract the situation**, supporting the victim by emphasising that they could not be blamed for the bullying (e.g., *kids* [can be cruel], [sad that] *people* [do this]).
- ***You are beautiful inside***: Support was expressed in the form of **compliments** for victims for their appearance (*beautiful*) or for surviving ([you became] *stronger*).
- ***They are disgusting***: Bullies were **criticised** (*disgusting*).
- ***I can relate to that***: **Empathy** was offered (*relate, know* [how you feel]).
- ***Never stop loving yourself***: General **advice** ([talk to someone who] *understands*, *ignore* [bullies], [be] *strong*, [tell] *teacher*).

Study 1: Conclusions

- UK female influencer channels are almost universally supportive spaces for discussions about bullying.
- Although not noted before in the academic literature, abstraction/generalisation is a widely used instinctive peer strategy to support victims by emphasizing that victims should not blame themselves.
- http://mozdeh.wlv.ac.uk/Bullying.html

Study 2: “My ADHD hellbrain”: A Twitter data science perspective on a behavioural disorder

Thelwall, M., Makita, M., Mas-Bleda, A., & Stuart, E. (2021). “My ADHD hellbrain”: A Twitter data science perspective on a behavioural disorder. *Journal of Data and Information Science*, 6(1). https://doi.org/10.2478/jdis-2021-0007

Study 2: Motivation

- Social media, including Twitter, may be used to publicly share experiences of medical conditions.
- Analysing tweets may give new insights into the patient perspective.
- Attention Deficit Hyperactivity Disorder (ADHD) is a common behavioral condition, affecting school and life.
  - Characterized by inattention and/or hyperactivity and impulsiveness.
- Tweets from people claiming ADHD may therefore give insights into the condition.

Study 2 Data: My ADHD tweets

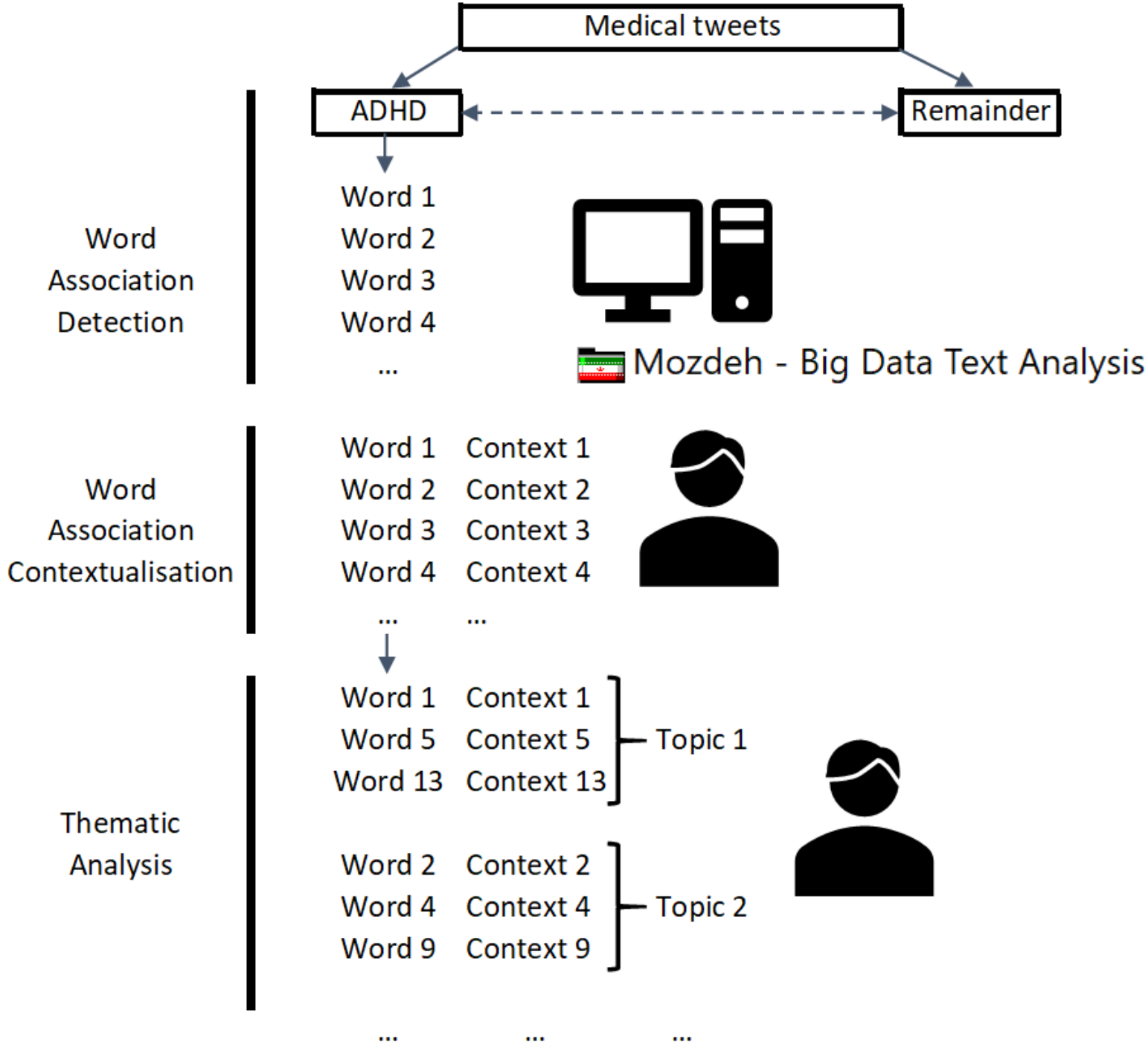
- Used query “my ADHD” to get tweets likely from people claiming to have ADHD.
- Comparator set of “my X” tweets for 99 common other medical conditions.
- Tweets collected every 15 minutes by free software Mozdeh July 2019 to February 2020.
  - 58,893 non-duplicate ADHD tweets.

Study 2 Method 1: Thematic analysis

- Three coders applied reflexive thematic analysis to a random set of 200 ADHD tweets.
- Objective: to identify main themes in ADHD tweeters' tweets about ADHD.
- Read tweets, tagged them with themes, organised themes, compared themes between coders, produced an agreed set of themes.

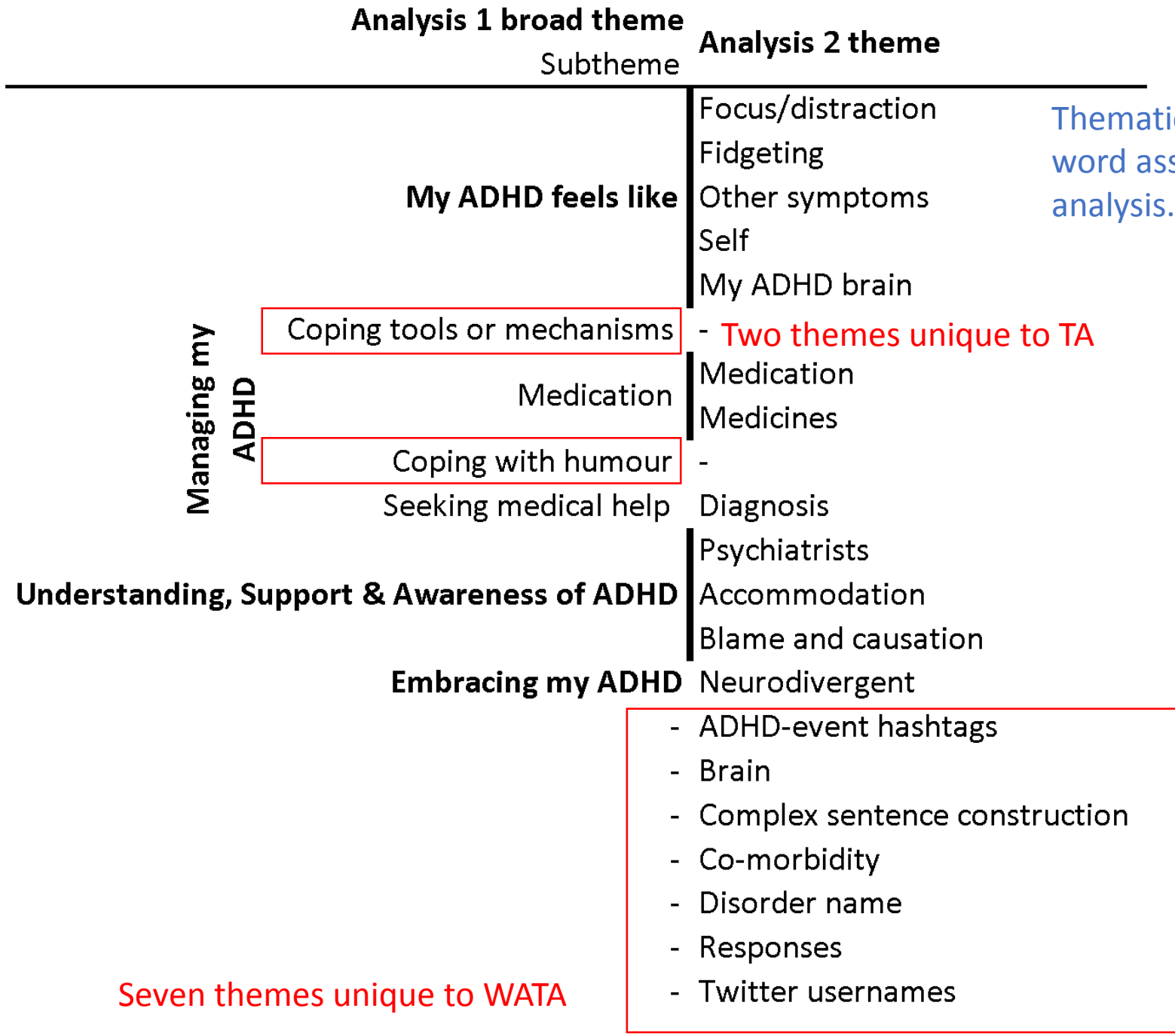
Study 2 Method 2: Word Association Thematic Analysis

- 200 words extracted that occurred in a higher percentage of ADHD tweets than other condition tweets.
- Thematic analysis applied to these words to detect themes discussed for ADHD more than for other conditions.


Medical tweets
ADHD
Remainder
Word 1
Word 2
Word 3
Word 4
...
Word Association Detection
Mozdeh - Big Data Text Analysis
Word 1 Context 1
Word 2 Context 2
Word 3 Context 3
Word 4 Context 4
... ...
Word Association Contextualisation
Word 1 Context 1
Word 5 Context 5
Word 13 Context 13
Topic 1
Word 2 Context 2
Word 4 Context 4
Word 9 Context 9
Topic 2
Thematic Analysis

Study 2 WATA themes (sample)

- **My ADHD brain**: The phrases “my ADHD brain” and “my ADHD hellbrain” distanced the tweeter from their actions.
- **Neurodivergent**, “I am starting to accept my neurodivergent status”: Used as a positive term.
- **Self**, “I hate myself sometimes”: Personal references (e.g., I, I’m) are more common, perhaps to explain ADHD-related behaviour.
- **Blame and causation**: Words expressing blame or causation (e.g. blame, because, so) are more common, due to the need to explain personal behaviour.


Analysis 1 broad theme
Subtheme
Analysis 2 theme
My ADHD feels like
Focus/distraction
Fidgeting
Other symptoms
Self
My ADHD brain
Managing my ADHD
Coping tools or mechanisms
- Two themes unique to TA
Medication
Medication
Medicines
Coping with humour
-
Seeking medical help
Diagnosis
Understanding, Support & Awareness of ADHD
Psychiatrists
Accommodation
Blame and causation
Embracing my ADHD
Neurodivergent
- ADHD-event hashtags
- Brain
- Complex sentence construction
- Co-morbidity
- Disorder name
- Responses
- Twitter usernames
Seven themes unique to WATA
Thematic analysis (left) vs. word association thematic analysis.

Study 2 Conclusions

- People with ADHD tweet to explain/justify symptoms.
- Use neurodivergence concept to be positive about ADHD.
- Both WATA and thematic analysis gave insights into ADHD perspective, at least as expressed on Twitter.
- The two methods give different information.
  - WATA can be more fine-grained and suggest new perspectives.

WATA Advantages and Disadvantages

- Advantages
  - A systematic method to turn an interest in a topic into a research study
  - Data collection is relatively straightforward
  - Some flexibility in method
- Disadvantages:
  - Mozdeh is Windows-only
  - Needs a relevant split in the data (e.g., new vs. old; male vs. female)
- Next steps
  - Find a topic that is interesting and tweeted about and try out WATA! Or
  - Recode WATA for Linux

References

- Thelwall, M. (2021). Can Twitter give insights into international differences in covid-19 vaccination? Eight countries' English tweets to 21 March 2021. *Profesional de la Información*. 30(3), e300311. https://doi.org/10.3145/epi.2021.may.11
- Thelwall, M., Makita, M., Mas-Bleda, A., & Stuart, E. (2021). "My ADHD hellbrain": A Twitter data science perspective on a behavioural disorder. *Journal of Data and Information Science*, 6(1). https://doi.org/10.2478/jdis-2021-0007
- Thelwall, M. & Cash, S. (2021). Bullying discussions in UK female influencers' YouTube comments. British Journal of Guidance and Counselling, 49(3), 480-493.
- Thelwall, M. (2021). Wor*d association thematic analysis: A social media text exploration strategy*. San Rafael, CA: Morgan & Claypool.