

# Word Association Thematic Analysis: Insight Discovery from the Social Web

Mike Thelwall <sup>a</sup>

*Statistical Cybermetrics and Research Evaluation Group, University of Wolverhampton,  
Wulfruna Street, Wolverhampton, U.K.*

**Keywords:** Word Association Thematic Analysis, Social Media Analysis, Twitter, Youtube.

**Abstract:** Billions of short messages are posted daily to the public social web. This gives opportunities for researchers to gain insights into the issues discussed, but extracting useful information can be challenging. On the one hand, the simplifying quantitative approaches for large scale analysis risk misinterpreting the patterns found because of the many different uses of the social web. On the other hand, small scale qualitative investigations may miss the big picture and ignore most of the data. This talk describes a mixed methods approach, word association thematic analysis, that attempts to gain the face validity of small-scale qualitative investigations with the power of large-scale pattern detection. The method leverages comparisons to identify sets of characteristic words for a topic, then applies thematic analysis to group these words into patterns according to the context in which they are used. The comparisons can be temporal (e.g., early vs. late tweets), topic-based (e.g., vaxxers vs. antivaxxers), or user-based (e.g., gender, location). The outcome of word association thematic analysis is a set of themes that characterise an issue in a social web site, supported by qualitative evaluations of the context of the words analysed and statistical tests for the validity of the differences identified.


## 1 INTRODUCTION

Market researchers and academics previously had to use slow, labour-intensive methods to test public opinion on issues such as politics, brands, and health concerns. These methods included surveys, interviews, focus groups, and user panels. Although these now have faster and cheaper web alternatives, they are now supplemented by datamining social media. For example, businesses can monitor attitudes towards their brands in real time by identifying and analysing relevant social media conversations. The same opportunities also exist for researchers, but the effective exploitation of social media data is conditional on the availability of acceptable analysis methods.

Typical problems for social media analysis compared to, for example, good quality surveys, is that the users of any social media site are a self-selected, biased subset of the population, their posts are influenced by the site and their friends and interests within the site, and the purpose of a post may

be unclear from its context. Whilst these problems are impossible to bypass for most purposes, social media analysis is still appropriate to provide quick insights into public opinion, if the limitations are understood, and to identify issues that researchers have not considered before.

There are two common approaches to social media analysis: small and large scale. Small scale studies have applied content analysis or thematic analysis to a sample of posts related to a topic to investigate what is discussed or how. Large scale studies have applied counting methods to map the structure of social media discussions (e.g., topic or user networks), to detect the main terms discussed (e.g., word clouds) or to mathematically model information diffusion and flow. This article discusses a new hybrid method, word association thematic analysis (WATA), which combines small scale analyses of individual posts with large scale selection of relevant posts to detect the key differences between two sets of posts. A focus on differences can bypass to some extent some of the sampling problems of

<sup>a</sup> <https://orcid.org/0000-0001-6065-205X>

social media data mentioned above. Identifying differences can be a specific aim of the project, such as to find gender differences in tweeting about a health condition and to discover how two contrasting political issues or brands are commented about on YouTube. It can also be a device to help explore a general topic. For example, finding differences between recent and older tweets about open-source software may help give insights into how people discuss it today.

## 2 TWITTER AND YOUTUBE

This article focuses on two sources of social media data, Twitter and YouTube, describing statistically informed methods to extract patterns from differences in comments/tweets within them for a given topic.

Twitter is an important source of news and information updates, which seems to be particularly used by professionals and in Western and English-speaking countries. A free academic API (<https://developer.twitter.com/en/products/twitter-api/academic-research>) seems to give access to all tweets ever sent, other than Twitter-identified spam, deleted and private tweets. This makes it an easy source of information about public opinion since Twitter became mainstream (it started in 2006). This is good for any topic that is extensively tweeted about, such as major news stories and academic events, but not good for more private issues, topics rarely tweeted about, or countries where Twitter is banned or unpopular.

According to the now defunct traffic monitoring website Alexa.com, YouTube has been one of the three most visited websites in the world for over a decade. It is enormously popular for watching a wide range of videos, such as for education, entertainment, and personal help (e.g., lifestyle YouTubers, DIY advice). There is also one comment for every thousand video views. Although these comments often seem to be very short (e.g., “great video!”), they collectively provide a corpus of public texts related to the topics of videos. They can be used to explore user reactions to topics that are well represented on YouTube. YouTube already gives extensive free access to its data so it is a good source of social web text for research. It is more technically difficult to collect a useful set of YouTube comments than a good set of tweets, however, because comments must be obtained indirectly via their videos and cannot be searched for directly. Moreover, searching for videos on a topic is not straightforward. This is perhaps the main reason that YouTube is researched far less often

than Twitter. The lack of content in most comments also does not help. This is unfortunate, given the much wider range of topics covered on YouTube and its much larger audience compared to Twitter.

## 3 WATA WITH MOZDEH

Word association thematic analysis is a method to detect themes reflecting textual differences between two sets of texts, usually from social media. For example, if the two sets were tweets about bullying from male and female tweeters, then the WATA goal would be to detect the main male-female gender difference themes in tweets about bullying. It is a multi-stage method that is supported by the free software Mozdeh ([mozdeh.wlv.ac.uk](http://mozdeh.wlv.ac.uk)). The early stages are mainly automated, and the later stages involve human reading of the texts. WATA works as follows.

### 3.1 Data Collection

Mozdeh can collect Tweets with the public API or Academic API and YouTube comments via the public API. The first step is to define the research topic and convert it into queries to match relevant tweets or YouTube videos for comments. Mozdeh also works with pre-collected texts from other sources (e.g., research article abstracts: Thelwall & Maflahi, 2015), which can be imported, bypassing the Twitter/YouTube stages.

For Twitter, the research topic must be translated either into a set of queries in the form of words and phrases that must be in the tweets, or a set of relevant users (e.g., Maltese politicians’ Twitter accounts). Unless the queries are defined by the research goal (e.g., a key hashtag: Potts & Radford, 2019), care must be taken to curate a set of relevant queries. These should be a (possibly large) set of words and phrases that have high precision (almost all the tweets containing them are relevant to the topic) and as high collective recall (total number of matches to the queries) as possible. It is essential to spend a day brainstorming and trying out different queries on Twitter.com to get a good set. Filtering out words and phrases generating too many false matches is particularly important or the rest of the method will be much more difficult.

For YouTube, comments can only be downloaded for specified videos, so the task is to identify a set of videos related to the research topic. This is easiest if the focus of the project can be defined in terms of YouTube users/channels because Mozdeh can find all

videos from those channels and download all their comments in one step. For example, a project to investigate the main news topics in India fed Mozdeh with the channel names of the five top Hindi news channels on YouTube (Deori et al., 2022). Influencers on YouTube have a large following, so their videos tend to attract many comments and they are a good and underused source of comments about aspects of daily lives for younger people. One study identified many UK-based female YouTube influencers and fed their channel names into Mozdeh to investigate discussions of bullying in their comments (Thelwall et al., 2022). Finding these influencers was difficult, however, and took several weeks of searching YouTube and news stories about UK influencers. If the YouTube topic can't be translated into YouTube users/channels, then YouTube.com must be manually searched to generate a list of relevant videos. It is only possible to automate this in Mozdeh for very distinctive topics (e.g., dance styles: Thelwall, 2018).

If the research topic was participant reactions to WEBIST conferences, then the Twitter queries might be all the official and unofficial WEBIST hashtags (e.g., #WEBIST, #WEBIST2021, but not the term WEBIST since a Twitter search suggests that half of the tweets using this word are irrelevant). For YouTube, there does not seem to be an official WEBIST channel, but a few conference videos could be found by searching the site (e.g., <https://www.youtube.com/watch?v=wXPR45dKPdc>), and a list could be built for entering into Mozdeh. In both cases there might not be enough data for an effective WATA because thousands of texts are needed for the methods to work well.

### 3.2 Word Association Detection (WAD)

Once the texts have been collected by Mozdeh or imported into it from another source, the next stage is for Mozdeh to split the texts into two sets (two sets and the remainder; or one key set with the remainder being set B) to start the comparisons (Figure 1). This split may be based on anything practical, including user gender (male vs. female), time (earlier years vs. later years), topic (topic A vs. topic B) users (set A vs. set B), popularity (more retweeted tweets vs. less retweeted tweets) or sentiment (positive texts vs. negative texts) for example. Mozdeh has a range of filters to allow this split to be specified. Texts not matching the two sets are ignored. For example, if comparing tweets from males and females then tweets would be ignored from nonbinary users or users for which a gender could not be determined.

Mozdeh has a word association detection button ('Mine Associations...') to identify words that occur more in one set than the other. It works by extracting all words from all texts and then identifying those that occur disproportionately often in one set compared to the other. For instance, if the word 'Valletta' occurred in 5% of one set but 1% of the other then Mozdeh would notice the difference. Mozdeh then lists the words occurring disproportionately in one set, using a statistical chi squared test of the difference between the proportions to arrange them in order of statistical significance. The most significant words tend to be those that occur the most often and with the largest percentage difference between the two sets. This list gives the first insights into systematic differences between the two sets of texts.

The statistical test used by Mozdeh is important because differences between the two sets of texts can occur by chance. An additional procedure is needed to ensure the statistical significance of the results due to the complication caused by conducting multiple simultaneous tests (the familywise error rate issue), and Mozdeh uses the Benjamini-Hochberg procedure for this (Benjamini & Hochberg, 1995), reporting statistical significance with stars in the results. Despite this precaution, the tests need to be treated with caution because they assume that the words in texts for a topic are independently generated, which is untrue because tweeters may be influenced by each other and by external events, such as news stories.

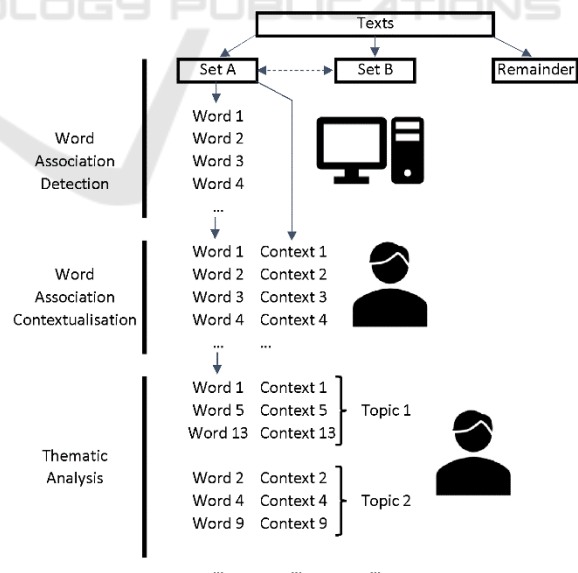


Figure 1: The main WATA stages after data collection (Thelwall, 2021).

For a WEBIST discussions on Twitter, the sets might be tweets expressing positive or negative sentiment about

the conference. The word lists produced would therefore be words occurring disproportionately often in positive or negative tweets. These would include direct sentiment words (e.g., amazing, wonderful, dull) and common sentiment targets (e.g., speaker, talk, venue, Malta, interruption, slides), the latter being the most useful.

### 3.3 Word Association Contextualisation (WAC)

The lists of words that occur in statistically significantly different proportions in the two sets may give some insights into the differences, and might be displayed in word clouds, tables or bar charts, but words are not enough. This is because the meaning of a word is affected by its context. The word ‘bank’ could mean river bank, financial bank, hillside or angle, for instance, and this would not be evident from its presence in a word cloud. The same is true to some extent for unambiguous words: ‘unexpected’ only has one meaning but only the context of the word in its tweet or comment would reveal what was unexpected. The next WATA stage is therefore contextualisation: finding out the context in which each statistically significant word was typically used (Figure 1).

The word association contextualisation (WAC) stage involves one or more people reading a random sample of texts from set A (or set B, whichever contains the highest proportion of the word) to identify its most common context. For example, the context of the word ‘unexpected’ within WEBIST tweets might be ‘unexpected praise’ because most tweets in set A containing the term unexpected mentioned being surprised when a stranger said that they really liked their conference paper. This is a content analysis type of task, which is a common social science research method (Neuendorf, 2017).

In general, reading 10-40 texts gives a good idea of the typical context of a word (if any) and allows it to be summarised in a sentence. This can be repeated for all the statistically significant texts, if there are not too many, or for a sample of a few hundred otherwise. This process is labour intensive but has the extra advantage that reading the texts helps to create a deeper understanding of the topic.

### 3.4 Thematic Analysis (TA)

The list of contexts for words occurring disproportionately often in set A or B is likely to have many overlaps and themes. For instance, the list might contain both ‘surprised’ and ‘unexpected’, with the same context, leading to redundancy when reporting the results. More generally, words may have

related contexts, even if they are not the same. For example, ‘question’ might have the WAC context of a question about a WEBIST presentation, and ‘praise’ may have the WAC context of praise for a WEBIST presentation, so the two contexts are related by both being reactions to a WEBIST presentation. Of course, if all tweets about WEBIST are reactions to presentations then this is an unhelpful generalisation but if they mostly discuss other topics (e.g., food, tourism, meeting old friends) then this might be useful to point out. The next WATA stage is identifying themes within the word contexts: Thematic analysis (TA). Again, this is a common social science research method (Clarke et al., 2015).

The thematic analysis stage involves attempting to cluster the word contexts from the WAC stage into themes, each containing related contexts. This involves trying to make appropriate generalisations about the contexts to make the themes. It is an iterative process without a correct solution, just reasonable sets of themes. One way of achieving this is to brainstorm for possible generalising themes for each context, then clustering the contexts together into the themes found as a starting point. Contexts that do not fit well into themes with other contexts might then be examined again for any commonalities in case they can fit into an even more general theme.

Since this process is subjective, ideally it would be conducted independently by multiple people, merging the results after discussion. This would help to improve the robustness of the method.

The result of WATA is a set of themes for set A and a set of themes for set B representing the main differences between them in the topic. The two sets are normally created independently but in the final stage the themes aligned as much as possible, if relevant, to aid interpretation.

## 4 WATA EXAMPLES

This section illustrates WATA with two examples of research project using it, for YouTube and Twitter.

### 4.1 Bullying on Youtube

A WATA study investigated how bullying was discussed in the comments to the videos of UK female lifestyle influencers on YouTube (Thelwall & Cash, 2021). The rationale for the study was that bullying is an important issue for young people and lifestyle influencers discuss personal issues. They are presumably watched by young people that have suffered from bullying so it is useful to understand



how they deal with the issue. Although there are bullying support websites, many victims might not seek help or talk about their experiences with friends, and so bullying-related discussions on general social media sites might provide anonymous indirect support or hostility, depending on the attitude of the influencer and their followers.

To start, YouTube was searched for UK-based female lifestyle vloggers with at least 20,000 subscribers, which was surprisingly difficult. 34 were identified and Mozdeh was used to download all 4.6 million comments on all their videos. Of course, few of the videos or comments were about bullying but about 8k were and the bullying set was defined as being the comments matching the following Mozdeh query of the comments (i.e., containing any of the words): *bullying bully bullied bullies cyberbully cyberbullied cyberbullies cyberbullying*. This was set A for WATA. Set B was the remaining comments so there was no 'Remainder' set for Figure 1.

For the WAD stage, only bullying-related words were analysed because non-bullying topics were irrelevant (i.e., set A but not set B words). This was achieved in Mozdeh by entering the above bullying query and clicking the 'Mine Associations...' button. This produced over 1000 bullying-related words but only the top 100 were analysed as a practical step because the WAD stage was time consuming and after 100 words, no new themes were emerging.

The WAC and TA stages created 12 bullying related themes, five describing bullying and seven expressing support for victims. The first set included the location of the bullying (school or online), its duration, how it happened and its long-term effects. The second themes included thanking and expressing support for victims, criticising bullies, praising the victims, expressing empathy, and offering general advice. One theme did not seem to have been mentioned before in the bullying literature: supporting victims by abstracting the situation to emphasise that they and their individual traits were not to blame for being bullied, despite this being the apparent focus of the bullying. This last point identifies a particular strength of WATA: its automatic identification of words at the start can help the researcher identify issues that they were not previously aware of.

## 4.2 ADHD Updates on Twitter

Attention Deficit Hyperactivity Disorder (ADHD) is a common behavioural disorder that can cause problems in some aspects of people's lives. A WATA study investigated how people with ADHD tweeted

about their condition in the hope of identifying new insights into the sufferer perspective compared to previous research using interviews or focus groups (Thelwall, et al., 2021). The rationale for the study was that Twitter is sometimes used to provide life updates and people with ADHD might tweet about things that they did not consider important enough to mention in interviews or surveys.

Although personal issues are not a natural topic for gathering tweets, ADHD is a common enough condition for it to be reasonably well represented on Twitter. Gathering tweets by people with ADHD is tricky, though, since many tweeters without ADHD can mention it (e.g., researchers, teachers, friends, family, organisations). This makes it difficult to generate high precision queries that target tweets by people with ADHD. The solution to this was to query the phrase, 'my ADHD', since checks at Twitter.com suggested that this phrase was almost exclusively used in tweets by people that appeared to be reporting about their own ADHD. This query had the additional advantage that tweets containing 'my ADHD' would not only probably be from people with ADHD but also probably be about something related to their ADHD. The latter point is important because of course people with ADHD may also tweet about the news, sport, computer games and anything else.

For the data collection stage, set A comprised tweets containing the phrase 'my ADHD' and set B included tweets containing the phrase 'my X', where X was any one of 99 other health conditions. It was important to compare against other self-declared health conditions to get insights that were specific to ADHD. Mozdeh collected 1m tweets for this, 59k of which were ADHD related.

For the WAD stage, only ADHD-related words were analysed because non-ADHD topics were irrelevant (i.e., set A but not set B words). This was achieved in Mozdeh by entering 'my ADHD' and clicking the 'Mine Associations...' button. This produced over 1000 ADHD-related words but only the top 200 were analysed because no new themes were emerging when this number was reached.

The thematic analysis stage identified 19 themes, although 4 were trivial (e.g., usernames). These included medication, focus/distraction, fidgeting, other symptoms, accommodations, diagnosis, psychiatrists, brain, 'my ADHD brain', neurodivergence, self, blame/causation, and comorbidities. For example, the theme 'my ADHD brain' was derived from WAD words including brain, hellbrain and ass.

Ass for the YouTube example, most themes reflected issues that were already known about in the

ADHD research literature, but some were new. In particular, the tendency of people with ADHD to discuss their brain as being a separate entity as an apparent coping strategy for themselves or communication with others did not seem to have been explicitly discussed in the academic literature before, although it is used in ADHD magazines.

This study also applied content analysis to the same data and compared the findings with those from WATA, showing that WATA was able to find themes not in the content analysis results. This is possible because WATA themes can be too rare to be identified by content analysis, despite being statistically significant.

## 5 CONCLUSIONS

Word association thematic analysis can generate insights into topics by leveraging differences within them to identify themes. It is a discovery-based method that can direct the researcher to previously unknown themes. It is supported by the software Mozdeh, which can help with all the automated stages, from data collection from Twitter and YouTube, to detecting word associations, and producing randomly sorted lists of texts to analyse. The main limitation of the method is that the topic must be discussed enough (thousands of posts) for it to identify meaningful themes.

More information about WATA can be found on the Mozdeh website, the WATA book (Thelwall, 2021) and in the methods sections of articles using it (e.g., with the phrase query ‘Word Association Thematic Analysis’ in Google Scholar).

## REFERENCES

- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), 289-300.
- Clarke, V., Braun, V., & Hayfield, N. (2015). Thematic analysis. *Qualitative psychology: A practical guide to Research Methods*, 222(2015), 248.
- Deori, M., Kumar, V., & Verma, M. K. (2022). What news sparks interest on YouTube? A study of news content uploaded by India's top five Hindi news networks. *Online Information Review*. <https://doi.org/10.1108/OIR-01-2022-0007>
- Neuendorf, K. A. (2017). *The content analysis guidebook*. Sage, Oxford, 2<sup>nd</sup> edition.
- Potts, G., & Radford, D. R. (2019). #Teeth&Tweets: the reach and reaction of an online social media oral health promotion campaign. *British Dental Journal*, 227(3), 217-222.
- Thelwall, M. (2018). Social media analytics for YouTube comments: Potential and limitations. *International Journal of Social Research Methodology*, 21(3), 303-316.
- Thelwall, M. (2021). *Word association thematic analysis: A social media text exploration strategy*. Morgan & Claypool, San Rafael, CA. <https://doi.org/10.2200/S01071ED1V01Y202012ICR072>
- Thelwall, M. & Cash, S. (2021). Bullying discussions in UK female influencers' YouTube comments. *British Journal of Guidance and Counselling*, 49(3), 480-493. <https://doi.org/10.1080/03069885.2021.1901263>
- Thelwall, M., & Maflahi, N. (2015). How important is computing technology for library and information science research? *Library & Information Science Research*, 37(1), 42-50.
- 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), 13-34.
- Thelwall, M., Stuart, E., Mas-Bleda, A., Makita, M., & Abdoli, M. (2022). I'm nervous about sharing this secret with you: YouTube influencers generate strong parasocial interactions by discussing personal issues. *Journal of Data and Information Science*, 7(2), 31-56.