Multi-modal Considerations for Social Media Discourse Analysis: A Specialised Corpus of Twitter Commentary on Working From Home

Authors:
Christopher Fitzgerald - Mary Immaculate College: christopher.fitzgerald@mic.ul.ie
Geraldine Mark – Cardiff University: markg2@cardiff.ac.uk
Anne O'Keeffe - Mary Immaculate College: anne.okeeffe@mic.ul.ie
Dawn Knight – Cardiff University: knightd5@cardiff.ac.uk
Justin McNamara - Mary Immaculate College: justin.mcnamara@mic.ul.ie
Svenja Adolphs – University of Nottingham: svenja.adolphs@nottingham.ac.uk
Leigh Clark – Swansea University: l.m.h.clark@swansea.ac.uk
Benjamin Cowan – University College Dublin: benjamin.cowan@ucd.ie
Tania Fahey Palma – University of Aberdeen: t.faheypalma@abdn.ac.uk
Fiona Farr – University of Limerick: fiona.farr@ul.ie
Sandrine Peraldi – University College Dublin: sandrine.peraldi@ucd.ie

Abstract

Social media discourse has evolved beyond merely textual communication into a more dynamic multi-modal communicative act with text supported by photographs, videos, emojis, gifs and other media. While social media discourse analysis facilitates rich investigation into public sentiment towards topical issues (Berber Sardinha 2022), there are challenges faced when compiling text-based corpora containing these multiple media. Extracting a holistic understanding of social media communication is reliant on these visual media to support the text-based discourse. Therefore, an analysis of the textual communication alone is a degraded platform from which to establish a rigorous interpretational framework. This chapter investigates a corpus of tweets extracted and compiled during a period of transition for workers experiencing a move from working from home to a return to the office workplace in the post-Covid environment (January-February 2022). While Twitter data-mining tools are attractive for their capacity to extract large datasets efficiently (Ruiz-Soler 2017), we argue that a smaller, specialised corpus of manually extracted tweets can provide a more principled and holistic analysis with consideration given to multiple media. Ethical barriers and

considerations regarding the use of this data are explored and an assessment of manual approaches to corpus construction and analysis of this data is provided.

1.0 Introduction

Social media, as defined by Kaplan and Haenlein (2010: 61) is 'a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content.' Posts and comments are open to further commentary, posting and reposting, creating what Wortham and Reyes (2021: 148) describe as a forum that is 'densely networked', bringing social media into the territory of conversation and dialogue between participants. Moreover, as found by Clarke (2022), in her multi-dimensional analysis of a large corpus of tweets, Twitter discourse often resembles spoken discourse, with contracted forms and an informal register seen as salient, providing researchers of Twitter data with a potential avenue for the analysis of speech-like data. The growth and mass sharing of this content through publicly accessible platforms has generated huge volumes of discourse which can be extracted or 'scraped' to be treated as linguistic data. Scraping tools have facilitated the automatic extraction of millions of words of social media discourse in a matter of seconds and at little or no cost. Further natural language processing tools can provide various metrics to analyse this data such as gauges of sentiment and frequency of terms which inform the researcher regarding the characteristics of such large datasets. These analyses provide researchers with a sense of how this relates back out to society and throws light on social perspectives on various matters, with researchers such as Zhang *et al.* (2022: 3) seeing social media as a 'new way to conceptualise public opinion'. While this process is attractive in terms of efficiency and capacity for big-data processing and analysis, we argue in this chapter that manual approaches to both data extraction and analysis have a number of benefits such as common-sense approaches to text selection and holistic

treatment of multimodal texts. To ensure that the multiple modes of content, characteristic of this type of discourse, are included (other than text alone), an intermodal approach (Zappavigna and Logi 2021) is undertaken to determine that which is constructed via the combination of text with other media, such as images.

This study was undertaken as a preliminary phase of the Irish Research Council and Arts and Humanities Research Council funded Interactional Variation Online (IVO) project (Knight *et al.* forthcoming). While the IVO project (see http://ivohub.com) as a whole seeks to linguistically analyse virtual meetings from a multi-modal perspective, this chapter focuses on a preliminary scoping phase for the wider project. This initial phase set out to provide the IVO researchers with an insight into public sentiment towards and experiences of working from home in the context of the COVID-19 pandemic, when many employees were requested to remain at home during lockdown measures.

Since its emergence in 2006, Twitter's growth in popularity has seen parallel levels of interest in mining its contents as a resource for both business research and development (Naeem *et al*. 2022) and academic research purposes (Zimmer and Proferes 2014; Rachunok *et al.* 2022). According to Twitter (as presented by oberlo.com), as of 2022, there are 229 million daily active users on Twitter, 59.2% of whom are aged between 25 and 49. There are 500 million tweets sent per day and the average time spent on Twitter is 3.39 minutes. A review of research carried out using Twitter by Karami *et al.* (2020: 67709) found that the most discussed topics were 'sentiment analysis, social network analysis, big data mining, topic modelling, and content analysis.' The studies based on linguistic analyses of Twitter data usually entail the construction of a corpus, defined by Carter and McCarthy (2006: 10) as 'a collection of texts, usually stored in computer-readable form.' Corpora typically represent a collection of a particular variety or use of language. Several existing corpora of tweets have been created for use by researchers. These include content specific corpora, such

as Saniei and Docel's (2022) corpus of tweets relating to disclosure of health information during the COVID-19 pandemic, time stratified corpora such as the Broad Twitter Corpus (Derczynski *et al.* 2016) and corpora made with the goal of gauging sentiment, such as the Moral Foundations Twitter Corpus (Hoover *et al.* 2020). The publication and distribution of such corpora present issues regarding ethical parameters (see Weller 2014) which we will return to later in this chapter. There is an attraction to engage with social media for research purposes due to its convenient accessibility and ease of collection. As Tefukci (2014: 505) notes, datasets may be created which take 'relatively little effort compared with traditional sociological methods'. This raises several questions. Is all of this content worthy of linguistic analysis? With such an abundance of content, how does one analyse the content in a principled way? How does a researcher approach social media data that is heavily reliant on extratextual media such as images and video? What are the ethical parameters to undertaking such analyses? We attempt to explore these issues critically to make the analysis of such data more accessible to future research.

This chapter explores these challenges and draws conclusions surrounding the benefits of integrating analytical tools with manual approaches to corpus construction and interpretation. While acknowledging the advantages of contemporary tools such as FireAnt and AntConc (described below), we highlight some limitations in dealing with a qualitative analysis of dynamic social media data. As Larsson *et al.* (2022) outline, recent research in corpus linguistics has tended towards statistical analyses facilitated by large datasets and technology-enhanced analytical tools at the expense of qualitative linguistic description, finding that in 2009 more corpus linguistics articles tend towards linguistic description over statistical reporting, while, in 2019, the reverse of this is the case, a trend which Larsson *et al.* *d*escribe as 'troubling' (p.154).

With parallels to the current study, Zhang *et al.* (2021) performed a sentiment analysis on a corpus of 1 million tweets related to remote working, extracted using Twitter scraping packages. They found that sentiment towards working from home is largely positive during the regular working week but becomes negatively oriented at weekends. While their analysis provides a sense of the sentiment expressed in their large corpus, the tweets analysed represent non-textual content as textual tags in square brackets that are non-descriptive such as [weblink] and [image]. From a quantitative point of view, Zhang et al's study provides an impressive template for automated data collection and analysis, but remains limited in its capacity for holistic representation of data. In contrast, this chapter proposes a more manual approach to Twitter data collection and analysis that can be used to complement such analyses as that of Zhang *et al.* (2021) which exploit the quantitative capabilities of contemporary tools.

2. Data and Methodology

As Egbert *et al.* (2022: 129) recommend, 'corpora should be designed and created to be as large as necessary to achieve precise parameter estimates, but no larger.' There is no one size fits all solution to corpus construction or size (Carter and McCarthy 2001) but a corpus must be representative of the language that is being investigated (Reppen, 2022). We adhere to these principles in constructing the current corpus by extracting tweets focusing on one subject over a set period of time. The singular focus on one subject (working from home) determines a minimal standard deviation in terms of content i.e., the likelihood of the content of tweets veering from the topic of the search term is low as the selection of tweets is determined by their containing this term. The tweets that make up the corpus used in this study were manually extracted between January and February 2022. Tweets were initially identified by searching for the phrase 'working from home'. Tweets were extracted with our focus phrase as a hashtag, as well as tweets which contained the phrase. Taking this

approach provides more instances of the focus phrase used within sentences and broader phrases rather than the phrase in isolation as is often the case with hashtags (see for example Giaxoglou and Spilioti 2020) . Initially, 100 tweets were extracted from 'top tweets' to get a baseline sense of the use of this phrase at a given point in time, and a further 50 tweets were extracted every week for four weeks from the 'latest tweets' tab to gain insights into trends associated with the use of this phrase over time and to sample a broader time period. 'Latest tweets' were selected rather than 'top tweets' to avoid any algorithmic influence determined by the user's preferences ('top tweets' are selected for accounts based on past searches, likes and retweets). This resulted in a corpus of 300 tweets, totalling 7,928 words as well as extratextual content that is presented and discussed later. Table 1 gives a breakdown of tweet numbers over the five-week time period of data collection.

	Top tweets	**Latest tweets**	**Latest tweets**	**Latest tweets**	**Latest tweets**
Date of tweet	Dec 28th 2021- Jan 6th 2022	12th December 2022	20th January 2022	27th January 2022	3rd February 2022
Number of tweets	100	50	50	50	50

Table 1: Numbers and dates of tweets extracted

Though this may be regarded as a small corpus, we will see that it nonetheless provides adequate substance for the aims of this study and aligns with Flowerdew's (2004: 21) parameters for a specialised corpus in that it is a small-scale corpus with a specific purpose, based on a single discourse type.

This manual approach was considered preferable over Twitter scraping tools for the following reasons:

1) The selection of tweets from individuals and not companies or organisations. Tweets from institutions rather than individuals were disregarded as these can be misleading and agenda driven (Hagen *et al.* 2020).
2) The selection of tweets authored by humans and not by automated profiles or 'bots'.
3) The identification of extratextual material such as images, gifs and videos and the ability to manually note this during the process of extraction.

None of the commentaries or threads following the posts used in this data have been included in the corpus (see reasons (1) to (4) below) but it is worth stating that comments on tweets are much less likely to contain extratextual material than the original posts. While we noted the number of comments that followed each 'head' tweet to gauge reaction, we did not include the comments in the corpus because:

1) they did not usually contain the search phrase.
2) they are typically not possible to interpret in isolation from the context of the original tweet, making analysis challenging.
3) they are often direct statements of agreement or disapproval and are not as rich as the original tweet.
4) while some tweets instigate many hundreds of responses, others receive none, thus it was determined that the inclusion of a string of comments on a popular tweet might skew the data.

The tweets were copied into an excel spreadsheet with notes collected for each tweet describing the date of the tweet, any hashtags used, the handle of the user, the number of responses, the number of retweets and the number of likes. There are instantly issues surrounding this information as Twitter is an evolving medium and the popularity of tweets builds over time, so notes on the amount of likes and responses of a tweet should always come with the caveat of this being at the time of extraction as newer tweets are likely to have

fewer signs of engagement. Descriptive information was also gathered regarding extratextual content such as images. Emojis within tweets were tagged using the automated textual descriptions that are built into excel such as 'smiling face with heart-shaped eyes' and 'face with tears of joy' which were used as proxies for these for ease of use with corpus software. Two tools were used to analyse this corpus: AntConc version 4.1.0 (Anthony 2022) is used to provide lists of frequently occurring items and patterns in the corpus and FireAnt version 2.2.0 (Anthony and Hardaker 2022) is used to isolate and list items within specific columns of the file. While FireAnt has functionality to extract social media data, it is employed here as means to focus on different aspects of the corpus which FireAnt lists in single columns, such as the text used in the tweets or hashtags, prior to looking at how other elements in the tweet, such as images or emojis, co-create meaning and sentiment. This approach facilitates the isolation of tweets containing extratextual material, the central focus of the analysis below.

Tweets were first categorised into those that contain and those do not contain extratextual material. Sentiment and meaning were looked at from a 'Common Sense' approach (Van Hees *et al.* 2018). This is described by Cambria *et al.* (2009: 253) as the knowledge of 'the basic relationships among words, concepts, phrases and thoughts', the kind of knowledge which Cambria *et al.* acknowledge is often beyond the capabilities of machines. This approach contrasts with other studies which have relied heavily on Natural Language Processing (NLP) tools to make such determinations and to prioritise human interaction with this discourse over analyses grounded in criteria determined by computer-based tools. As Van Hee *et al.* (2018: 794) point out, 'This connotative knowledge, or typical sentiment related to real-world concepts, comes naturally to most people, but is far from trivial for computers.' This approach is also necessary to determine sentiment and meaning based on extratextual information, which NLP tools fall short of achieving; as Reddy and Agarwal (2021: 526) state, 'most people have a natural sense of common sense and

connotative knowledge, machines, however, struggle to execute tasks effectively that require extra-textual/contextual information.' Highlighting the disparity between manual sentiment categorisation and that afforded by sentiment analysis tools, Furini and Montanegro (2018) note that sentiment analysis of tweets by people is easy, but challenging for computers, though the former is much more time-consuming than the latter. Their study took a novel approach to bypassing this by gamifying sentiment analysis to make the process of sentiment categorisation enjoyable for people to participate in and thus, facilitating the capacity to undertake sentiment analysis of a large set of tweets by humans in a way that is entertaining. This offsets the time-consuming and often arduous task of manual sentiment categorisation of large datasets. We adhere to O'Halloran's (2014: 1) view of integrating multiple modes into our analysis, suggesting that the 'analysis and interpretation of language use is contextualized in conjunction with other semiotic resources which are simultaneously used for the construction of meaning.' As will be seen below, the meaning of tweets is often determined by the accompanying visual media which therefore needs to be considered in conjunction with the text to determine both meaning and sentiment.

As Zhang *et al.* (2021) found, the social and political context of the time when the tweets were extracted influenced the content. This was also seen as an influence in the current study when topical content of the time influenced the tweets, such as controversies surrounding the activities of residents of 10 Downing Street during the pandemic (directly referred to in 6 tweets). In addition, changes in working from home regulations arose as a prominent influence on the content of tweets regarding working from home.

Before continuing to the analysis of this data, we first turn to the ethical considerations in the treatment of such data, as this is an evolving and complex topic which warrants discussion.

3. Social media discourse and ethical considerations

Throughout the process described in section 2, ethical considerations frequently arose. Best practice in ethical research has traditionally stated that media and documentation in the public domain is accessible for research purposes without further consent, despite participants often not being aware that their posts may be used for research (Weinberg and Gordan 2015). Indeed, the question of whether researchers dealing with thousands of tweets regard their research as dealing with participants at all is raised in the British Association for Applied Linguistics' Recommendations in Good Practice in Applied Linguistics (2021: 9). The suggestion that the very existence of information in the public domain inherently grants researchers the warrant to access and utilise it has been questioned. This is partly due to the degree of user awareness about what being in the public domain means, since as Kern *et al.* (2016: 15) state, 'It is questionable what users understand public to mean'. Bolander and Locher (2014) also contend that definitions of public and private have changed and suggest that ethical decision-making should not rely solely on this dichotomy. Seeking out permission, from both the users (as found by Ahmed *et al.* 2017) and the platform, can also prove difficult, with linguistic researchers such as Dijkstra *et al.* (2021) finding Twitter unresponsive to requests for permission to use tweets for research purposes. It is for these reasons that measures were taken to ensure that this chapter does not include any examples from the corpus that may be traceable to an author or Twitter profile. While we do show images for the purpose of illustration and exemplification of results, we apply the strategies shown by Chen *et al.* (2021) to 'minimize the risk of privacy erosion' including anonymising users' accounts, not showing IDs and profile images and paraphrasing text. In cases where tweets or screenshots of tweets are used as examples in this chapter, we have made direct requests to authors seeking permission for reproduction. Despite proposals and calls for greater oversight and regulation of the use of social media content for research purposes

(Hayden 2013), there remains a degree of ambiguity surrounding the issue (Clyne *et al.* 2018).

Ethical determinations should be influenced by the degree to which authors are fully represented in shared work. Likewise, it is an issue of ethical probity to ensure that authors or a population sample are not misrepresented. This may arise when content which is key to the meaning and interpretation of a message is removed from a tweet, as often happens when social media content contains extratextual media that is not readable by analytical tools. Though Artificial Intelligence and Natural Language Processing tools are becoming increasingly sophisticated at determining non-literal meanings and irony (Barbieri and Saggion 2014), there is still some way to go before such tools will be able to fully unpack the nuances of non-literal communication, especially that which relies on multi-modal content. Features which are specific to a particular platform also contribute significantly to the message conveyed on this media. For example, when analysing tweets, it would be remiss to neglect the significance of hashtags as highly functional in conveying irony and the core message of tweets (Van Hee *et al.* 2018). It is apparent from reviewing studies with Twitter as their core data via the use of scraping tools that Twitter users are not always fully represented in such studies since the tools used can only extract and analyse a degraded from of the original content. Despite not being identified, it is the responsibility of the researchers to adequately represent the participants who have created their data. Thus, this chapter argues that the optimal mode of analysis to determine the semantic properties of Twitter data is a manual approach that fully represents the data by incorporating all elements of content.

4. Results

Of the 300 tweets collected, 210 (70%) contained text only and 90 (30%) included other media . Table 2 presents the breakdown of these tweets that contain extratextual content. The table shows that of the 30% containing extratextual media, 60% (54) are photographs

(including selfies) and the remainder are gifs, links to articles and websites, short videos and screenshots or other digital images (including memes).

Total tweets	Containing extra media	Tweets containing photographs	Gif	Link	Video	Screenshot/digital image
300	30% (90)	18% (54)	4.7% (14)	3% (9)	2.4% (7)	2% (6)

Table 2: Extratextual content of corpus

4.1 Themes

In classifying tweets into themes, it is firstly evident that two categories emerge: tweets either refer to personal commentary regarding working from home, or present a stance tied to social enforcement or social effects of working from home from a broader perspective. The first category typically includes content related to one's direct surroundings such as their home office set-up, for example,

(1)

After two years working from home I finally have a setup with a really good mic which cancels background noise well

while the second category often points to issues of social equality as a result of working from home, for example,

(2)

So, adults are expected to work from home 'for at least another 3 weeks' but kids are expected to go back to school this week? Is that right?

As the figure 1 shows, 252 of the tweets (84%) comment on personal experiences of working from home while the remaining 48 tweets (16%) comment on societal issues regarding working from home. As can be seen by comparing the orange portions of the bars, the tweets that offer social commentary are less likely to contain media in addition to the textual content. 7 of the 48 (16%) tweets that offer social commentary contain extratextual media in comparison with 83 of the 252 (33%) of the tweets that commenting on working from home.

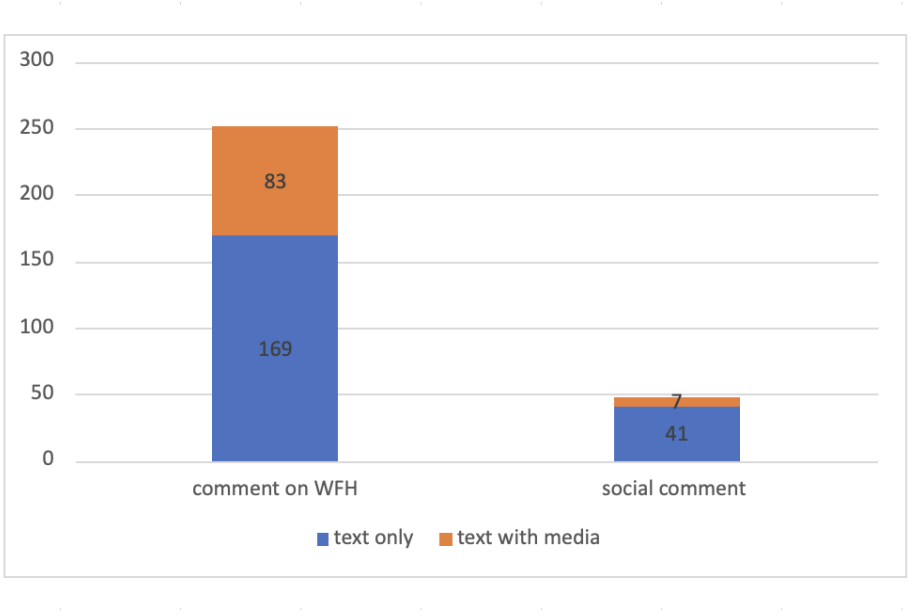


Figure 1: Broad thematic categories of tweets

Further themes within these broad categories reflect what Twitter users associate with working from home. Based on both textual and other content, the tweets are categorised into the sub-themes illustrated in figure 2.

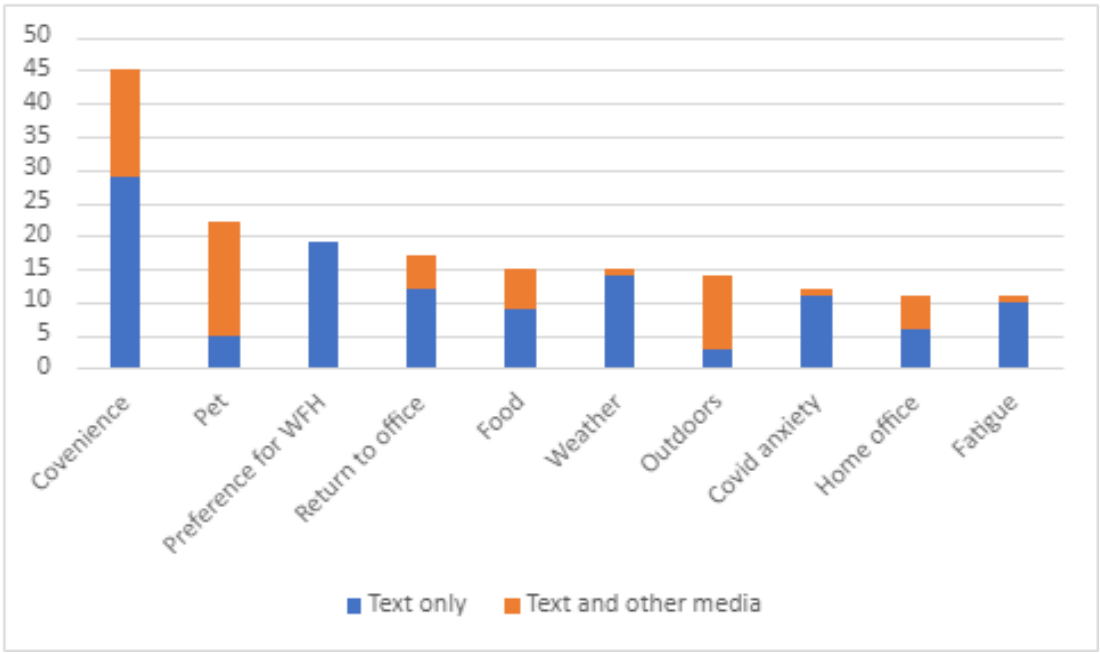


Figure 2 Sub-themes of tweets

The most prominent sub-theme is the expression of the convenience afforded by working from home. Many of the tweets which refer to the authors' direct surroundings include extra content such as images of what is referred to in the text, such as pets, food, outdoor environments and home office set-ups. As figure 2 shows, the topic of a tweet is a predictor of the likelihood of it containing extra media. For example, the majority of tweets referring to pets (76%) contain an image of the pet, while no tweets that express direct preference for working from home contain extra content.

4.2 Polarity of opinion and nuances of sentiment

As is often noted regarding social media discourse and particularly with Twitter, content often expresses extreme opinions and is seen as a gauge of polar perspectives (Yaqub *et al.* 2018). This is one of the defining characteristics of Twitter discourse and both attracts and repels users from the platform, a feature which may be more emphasised by the addition of a dislike button, which was tested in 2021, though largely rejected for its potential to heighten the 'toxicity' of Twitter discourse (https://www.irishtimes.com/business/technology/twitter-

begins-testing-dislike-button-but-there-s-a-catch-1.4627848). While the polarity of Twitter discourse adds to controversies surrounding its contribution to public discourse more broadly, this also makes it somewhat easier to categorise in terms of positive/negative sentiment as these are often expressed explicitly, either through textual or other means. Tweets that mention 'working from home' confirm this view of Twitter discourse. For example, one tweet describes working from home as 'hell', while another describes it as 'the best'. The most frequently-occurring verb preceding the search phrase, after various forms of *to be* is 'love'. In some instances, this can be misleading and may lead to mischaracterisation by sentiment analysis tools, which may regard the following as positive by not interpreting the irony embedded in the tweet:

(3)

gotta love working from home when your internet provider is a complete dumbass and does maintenance during working hours.

An analysis of clusters using AntConc software surrounding the search phrase at first reveals a sense of malaise towards working from home at the time of tweet extraction (December 2021-February 2022). The most frequent three-word phrase, apart from the search phrase, is 'I'm tired of'. This phrase occurs eleven times in the corpus and ostensibly points towards a negative attitude towards 'working from home'. Upon closer analysis, with the aid of FireAnt to observe individual tweets, it is revealed that nine of these occurrences appear in one tweet:

(4)

I'm tired.

I'm tired of those on the far left

I'm tired of those on the far right

I'm tired of Covid

I'm tired of working from home

I'm tired of eco doom & gloomers

I'm tired of anti-vaxxers

I'm tired of news

I'm tired of social media

I just want quiet.

Rather than calculating each of these expressions as one person's negative sentiment, a sentiment analysis tool such as LIWC-22 (Boyd *et al.* 2022) will add these to the overall metric of sentiment in a corpus and skew the resultant calculation. Though sentiment analysis tools such as LIWC-22 and others (see review by Al-mashhadani *et al.* 2022) provide researchers with a broad overview of sentiment, and they are becoming more sophisticated (Saif *et al.* 2016), they remain limited in their capacity to accurately gauge sentiment beyond the single word level (Mohammad 2017). This is further complicated in a study such as this which often relies on extratextual content to determine sentiment. As Van Atteveldt *et al.* (2021: 134) conclude in their assessment of automated sentiment analysis tools compared to human analysis, such tools 'do not always perform sufficiently'.

Models of automated classification of multi-modal social media data on emotional scales (Duong *et al.* 2017) and sentiment alignment (Graesser *et al.* 2017) have been developed, but remain inconclusive regarding full accuracy of interpretation. Of course, human interpretation of sentiment is challenging and made more complex by the addition of extra media such as images which may be ambiguous, ironic, sarcastic or bound to cultural references that may only be accurately interpreted by those who fully understand their source context. These challenges are analogous to those faced by sentiment analysis more broadly

(see Saura *et al.* 2022) and can be offset by certain categorisation criteria set by predetermined frameworks. The classification of sentiment used in this study is that used previously by Chen *et al.* (2012). In their classification of tweets, Chen *et al.* use four classifications; positive, negative, neutral and objective.

For this study, the objective classification is used in instances where no sentiment towards working from home is discernible in the tweet, for example:

(5)

Ireland has just made working from home compulsory in response to the surge in Omicron cases.

The neutral classification suggests an expression of sentiment that is neither positive nor negative or where both positive and negative sentiments are expressed in a tweet, for example:

(6)

I don't like working from home. But one piece of joy that I get from it; Toto relaxing by my side in front of the fireplace (accompanied by three heart emoticons and a picture of a cat).

In contrast to Chen *et al.,* who categorise sentiment based on sentiment expressions or root words, this study includes images or other media as sentiment-baring items within the tweets analysed. The sentiment towards working from home of tweets is thus categorised using Chen *et al.'*s parameters, but utilises the entire content of the tweets to determine this. While this system provides a framework for categorising individual tweets, it does not, like other sentiment classification systems, determine the extent to which tweets align with positive/negative sentiment or extremes on this cline. For example, 'working from home is the best thing ever' will result in the same classification as 'I like working from home'.

In applying this categorisation to our corpus of 300 tweets that mention 'working from home', we determined the sentiment classifications outlined in table 3.

Positive	Negative	Neutral	Objective
27.33% (82)	11.66% (35)	22.33% (67)	38.66% (116)

Table 3 Sentiment categorisation of tweets

The relatively high percentage of objective sentiment is due to many tweets mentioning working from home but not expressing a stance towards this. While sentiment may be expressed towards other topics within the tweet, such as politicians having garden parties during lockdowns, if the sentiment expressed was not directed towards working from home, then this was categorised as objective.

While Zhang *et al.'*s (2013) analysis included other terms related to working from home (such as 'remote work'), the results in this study (table 3) show similarities in that sentiment towards working from home emerged as more positive than negative. Table 3 illustrates that there were more than double the tweets showing positive sentiment (82) compared to negative sentiment (35), while Zhang *et al.'*s results find that 'average sentiment was slightly positive' (2013: 800). One factor in accounting for the disparity between positive and negative in the results in our study is the inclusion of content that is not text. As we will show, this extratextual content often contributes to the construction of meaning and sentiment - of the 90 tweets that contain such material, 32 of these are categorised as positive, while 9 are negative.

4.3 Intermodal considerations

In addition to the many discourse features marking social media discourse as different to face-to-face communication (Blommaert 2018), tweets contain a multitude of extratextual content such as images, gifs, emojis or videos which add to, support or replace text. Though

systems of image extraction and tagging have been explored (Safa *et al.* 2022), these remain limited in their capacity to determine meaning in combination with the textual content of tweets. Extracting the text from the extratextual content causes at least some degradation of the data and at worst a complete removal of the communicative purpose of the tweet. For example, our corpus is composed of several tweets with text that is just the search phrase 'working from home' accompanied by an image. While the search phrase in isolation provides the researcher with no gauge of an author's sentiment and the overall message remains ambiguous, an accompanying image of, for example, a laptop on a table with a tropical beach landscape in the background imparts a strong message to a human reader. In many cases, the text provided in tweets describes the content of images they accompany, performing a pragmatic function seen by Dainas and Herring (2021) as common in social media contexts where images echo or repeat the textual content of the message. For example, a tweet with an image of a landscape photograph taken out the author's window is accompanied by the following text:

(7)

incredible light and weather happening as I sit gazing out of the window (working from home)

While the textual content makes sense in isolation, its combination with the image places it in a supplementary role. Textual content is often used to make reference to an image. In such cases, the text content only makes full sense when the referent is known via the accompanying image. For example, the following text is accompanied by a photograph of the author with a cat:

(8)

Working from home affords genuine periods of joy when this one comes over for a snuggle.

There are instances in the corpus of clause completion via extra media. For example, the following text remains incomplete without the accompanying gif of an animated character dancing:

(9)

60% of my colleagues now caught covid within the last week. I'm working from home since late November. Dodging covid like

While extra media are not necessary for the comprehensibility of some messages, images are often used to provide evidence of what is expressed in the text. For example, a photograph of a beer is used in the message (10) as a means of providing evidence for the content of the text:

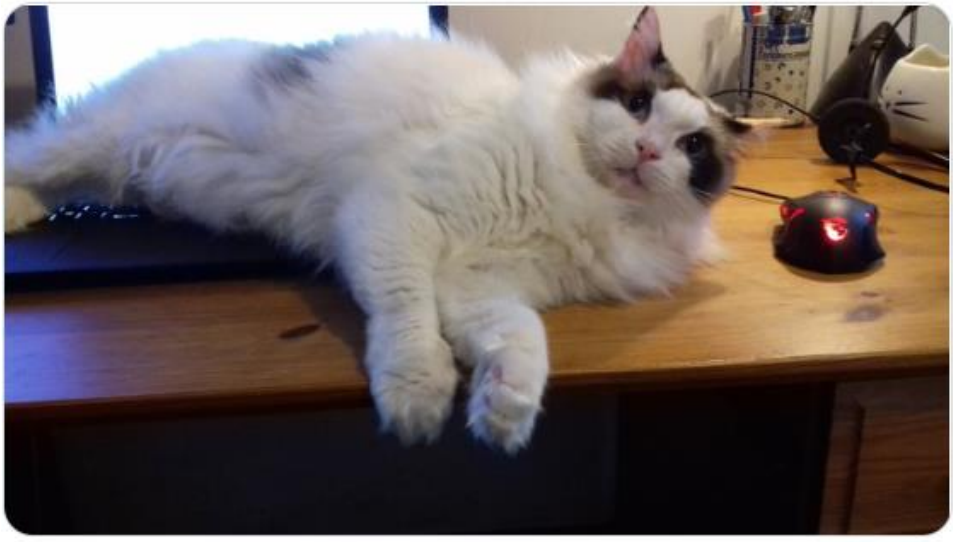
(10)

Just sat in my garden[office extension] working from home and getting pissed.

Textual content may make sense in isolation, as in tweet (11), but the tone of the message is carried by the photograph of the cat as the proposed 'technical issue' of the text:

(11)

working from home. Meant to be doing stress awareness training but having technical issues



As Macken-Horarik (2004: 5) states, ‘Attending to the multimodal poses particular challenges for discourse analysts who have worked primarily with verbal texts.’ In the case of Twitter analysis, these challenges have often been circumvented by neglecting non-textual content. By extracting and subsequently analysing text only, the data is likely to undergo semiotic impoverishment in the Peircean (in Bateman 2018) sense, that is, that signs and symbols contribute to meaning and to extricate them from the content is to degrade the discourse, and with a potential loss of semantic richness (Mautner 2016). In the case of Twitter discourse, meaning is often determined by the non-textual content. For example, the sentiment of the tweet (12) is difficult to discern without the observation of both the two-heart emoji and the photographs of a home-made meal that accompany the text:

(12)

Though the study of multi-modality in largely text-based material is not new (Stockl 2004) and should not be regarded as solely the dominion of social media discourse, this domain is one that is particularly reliant on images and other media to convey meaning. This relates particularly to Instagram and TikTok which are image and video-based platforms. A cursory observation of any Twitter feed will highlight the fact that the semantic properties of many tweets are reliant on extratextual content. For example, the sarcasm of the tweet (13) would not be detectable without the accompanying photograph and the cultural knowledge that it is a picture of Formula 1 car driver for Ferrari Charles le Clerc:

(13)

5. Conclusion

As social media usage has evolved, so too have the use and functions of visual content (Hasyim 2019). While Twitter consists of text as a key mode, other platforms such as Snapchat, Instagram and TikTok, which, according to the Pew Research Center (Auxier and Anderson 2021) are used far more by adults under 30 years old, are reliant on images and video (Brunner and Diemer 2019). This necessitates a reimagining of how to approach this platform from a research perspective and suggests the limitations of scraping tools that only extract text, and highlights a need to adopt a more semiotic approach such as the framework set out by Poulsen and Kvåle (2018).

The temptation to pursue these bigdatasets should not come at the expense of misrepresentation of data due to a neglect of media that may be best approached manually to provide accurate interpretation. Smaller samples that are representative of largerdatasets can provide the researcher with a holistic representation of an author's intended meaning. This offsets the increased reliance on statistical analyses rather than linguistic description which, as seen at the outset of this chapter, Larsson *et al.* (2022) have described as the trend in

corpus linguistics over the last decade. Removing content from its context in order to create a corpus has raised the question of the degradation of the authenticity of such corpora due to this displacement (see Mishan 2004). With regards to Twitter data, this is often compounded by the neglect of certain content due to it not being textual and easily treatable with scraping and corpus tools. We have provided examples of tweets whose semantic and sentimental properties may be misconstrued, without interpretation based on both visual and textual content. In the absence of tools which can analyse the combined semantic properties of images and extratextual material with text, there is a rationale for reverting to human, common sense analysis. This approach adheres to Leech's position that successful analysis is dependent on 'a division of labour between the corpus and the human mind' (1991: 15). This may necessitate the prioritisation of holistic analysis over large data samples, or the integration of both of these approaches to ensure the objective, quantitative rigour of NLP tools applied to largedatasets is met with the accuracy of semantic and sentiment interpretation of multi-modal data afforded by human interaction with smalldatasets.

Both the promise and pitfalls of using web archives will continue to be a point of discussion of humanities researchers from disparate fields including politics (Bastos and Mercea 2016), history (Milligan 2016) and linguistics (Sardinha 2022). This chapter has argued for the advantages of integrating rich platforms, such as Twitter that have the potential to provide an abundance of data, with more manual approaches to data treatment and analysis. This will allow us to arrive at interpretative conclusions that are less reliant on technological support and more dependent on a linguistic description that accounts for all components of the message. Though AI interpretive tools are advancing rapidly and are coming towards the capacity to derive accurate meaning from images, limitations remain evident in their ability to facilitate accurate readings of meaning or semantic from the combination of images with text, especially in cases where irony or sarcasm are latent. Until

these tools deliver this capacity, suggestions regarding accuracy of results should remain tentative.

References

Ahmed, W., Bath, P.A. and Demartini, G. (2017), "Using Twitter as a Data Source: An Overview of Ethical, Legal, and Methodological Challenges", Woodfield, K. (Ed.) The Ethics of Online Research (Advances in Research Ethics and Integrity, Vol. 2), Emerald Publishing Limited, Bingley, 79-107. https://doi.org/10.1108/S2398-601820180000002004

Al-mashhadani, M. I., Hussein, K. M., & Khudir, E. T. (2022), Sentiment analysis using optimized feature sets in different facebook/twitter dataset domains using big data. Iraqi Journal for Computer Science and Mathematics, 3(1), 64-70. DOI: https://doi.org/10.52866/ijcsm.2022.01.01.007

Anthony, L. (2022), AntConc (Version 4.1.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Anthony, L. and Hardaker, C. (2022), FireAnt (Version 2.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Auxier, B., & Anderson, M. (2021), Social media use in 2021. Pew Research Center, 1, 1-4.

Barbieri, F., & Saggion, H. (2014), Modelling irony in twitter. In Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics, 56-64.

Bastos, M. T., & Mercea, D. (2016), Serial activists: Political Twitter beyond influentials and the twittertariat. New Media & Society, 18(10): 2359-2378.

Bateman, J.A. (2018), 'Peircean Semiotics and Multimodality: Towards a New Synthesis' *Multimodal Communication*, 7(1): 20170021. https://doi.org/10.1515/mc-2017-0021

Blommaert, J. (2018), *Durkheim and the Internet: On sociolinguistics and the sociological imagination*, London: Bloomsbury Publishing.

Bolander, B., & Locher, M. A. (2014), 'Doing sociolinguistic research on computer-mediated data: A review of four methodological issues', *Discourse, Context & Media,* 3: 14-26. https://doi.org/10.1016/j.dcm.2013.10.004

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022), 'The development and psychometric properties of LIWC-22', Austin, TX: University of Texas at Austin. https://www.liwc.app

British Association for Applied Linguistics (2021), 'Recommendations on good practice in applied linguistics', *British Association for Applied Linguistics* https://www.baal.org.uk/wp-content/uploads/2021/03/BAAL-Good-Practice-Guidelines-2021.pdf

Brunner, M. L., & Diemer, S. (2019), 'Meaning negotiation and customer engagement in a digital BELF setting: A study of Instagram company interactions', *Iperstoria,* (13).

Cambria, E., Hussain, A., Havasi, C., & Eckl, C. (2009), 'Common sense computing: From the society of mind to digital intuition and beyond', *European Workshop on Biometrics and Identity Management,* 252-259, Berlin: Springer.

Carter, R. and McCarthy, M. (2006), *Cambridge grammar of English: a comprehensive guide: spoken and written English grammar and usage,* Cambridge: Cambridge University Press.

Chen, L., Wang, W., Nagarajan, M., Wang, S., & Sheth, A. (2012), 'Extracting diverse sentiment expressions with target-dependent polarity from twitter', In *Proceedings of the International AAAI Conference on Web and Social Media* 6 (1): 50-57.

Chen, Y., Sherren, K., Smit, M., & Lee, K. Y. (2021), 'Using social media images as data in social science research', *New Media & Society*. https://doi.org/10.1177/14614448211038761

Clarke, I. (2022), 'A Multi-Dimensional Analysis of English Tweets', *Language and Literature,* https://doi.org/10.1177/09639470221090369

Clyne, W., Pezaro, S., Deeny, K., & Kneafsey, R. (2018), 'Using social media to generate and collect primary data: The# ShowsWorkplaceCompassion Twitter research campaign', *JMIR public health and surveillance,* 4(2): 7686. https://doi.org10.2196/publichealth.7686

Dainas, A. R., & Herring, S. C. (2021), 'Interpreting emoji pragmatics. Approaches to Internet Pragmatics', in C. Xie, F. Yus, H. Haberland (eds), *Approaches to Internet Pragmatics: theory and practice,* Amsterdam/Philadelphia: John Benjamins, 107-144

Derczynski, L., Bontcheva, K., & Roberts, I. (2016), 'Broad twitter corpus: A diverse named entity recognition resource', *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers,* 1169-1179.

Dijkstra J, Heeringa W, Jongbloed-Faber L and Van de Velde H (2021), 'Using Twitter Data for the Study of Language Change in Low-Resource Languages. A Panel Study of Relative Pronouns in Frisian', *Artif. Intell.* 4:644554. https://doi.org/10.3389/frai.2021.644554

Egbert, J., Biber, D., & Gray, B. (2022), *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness.* Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316584880

Flowerdew, L. (2004), 'The Argument for Using English Specialized Corpora to Understand Academic and Professional Settings', in U. Connor and T. A. Upton (eds) *Discourse in the Professions,* Amsterdam: John Benjamins, 11– 33.

Furini, M., & Montangero, M. (2018), 'Sentiment analysis and twitter: a game proposal', *Personal and Ubiquitous Computing*, 22(4), 771-785. https://doi.org/10.1007/s00779-018-1142-5

Giaxoglou, K., & Spilioti, T. (2020), 'The shared story of# JeSuisAylan on Twitter: story participation and stancetaking in visual small stories', *Pragmatics,* 30(2), 277-302. https://doi.org/10.1075/prag.18057.gia

Hagen, L., Neely, S., Keller, T. E., Scharf, R., & Vasquez, F. E. (2020), 'Rise of the machines? Examining the influence of social bots on a political discussion network', *Social Science Computer Review,* https://doi.org/10.1177/0894439320908190

Hasyim, M. (2019), 'Linguistic functions of emoji in social media communication', *Opcion*, 35.

Hayden, E. C. (2013), 'Guidance issued for U.S. internet research', *Nature,* 496, 411. http://dx.doi.org/10.1038/496411a

Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., & Dehghani, M. (2020), 'Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment', *Social Psychological and Personality Science,* 11(8): 1057-1071.

Kaplan, A. M., & Haenlein, M. (2010), 'Users of the world, unite! The challenges and opportunities of Social Media', *Business horizons*, 53(1): 59-68. https://doi.org/10.1016/j.bushor.2009.09.003

Karami, A., Lundy, M., Webb, F., & Dwivedi, Y. K. (2020), 'Twitter and research: A systematic literature review through text mining', *IEEE Access*, 8, 67698-67717. https://doi.org/10.1109/ACCESS.2020.2983656

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016), 'Gaining insights from social media language: Methodologies and challenges', *Psychological methods*, 21(4), 507. https://doi.org/10.1037/met0000091

Koester , A. (2021), ‘Building Small Specialised Corpora’ in McCarthy, M., and O’Keeffe, A. (eds), *The Routledge Handbook of Corpus Linguistics* London: Routledge, 48-61. https://doi.org/10.4324/9780367076399-5

Larsson, T., Egbert, J., & Biber, D. (2022), ‘On the status of statistical reporting versus linguistic description in corpus linguistics: A ten-year perspective’, *Corpora,* 17(1): 137-157. https://doi.org/10.3366/cor.2022.0238

Leech, G. (1991), ‘The state of the art in corpus linguistics’ in K. Aijmer and B. Altenberg (eds) *English corpus linguistics: Studies in honour of Jan Svartvik*, 8–29 London: Longman, 8–29.

Macken-Horarik, M. (2004), ‘Interacting with the multimodal text: reflections on image and verbiage in Art Express’, *Visual communication*, 3(1): 5-26. https://doi.org/10.1177/1470357204039596

Mautner, G. (2016), 'Checks and balances: How corpus linguistics can contribute to CDA', in Wodak, R., and Meyer, M. (eds) *Methods of critical discourse studies*, London: Sage, 155-180.

Milligan, I. (2016), ‘Lost in the infinite archive: The promise and pitfalls of web archives’, *International Journal of Humanities and Arts Computing*, 10(1): 78-94. https://doi.org/10.3366/ijhac.2016.0161

Mishan, F. (2004), ‘Authenticating Corpora for Language Learning: A Problem and its Resolution’, *ETL Journal* 58(3): 219-227.

Mohammad, S. M. (2017), ‘Challenges in sentiment analysis. In Cham. E. *et al.* (eds.), *A practical guide to sentiment analysis*, 61-83. https://doi.org/10.1007/978-3-319-55394-8_4

Naeem, M., Jamal, T., Diaz-Martinez, J., Butt, S.A., Montesano, N., Tariq, M.I., De-la-Hoz-Franco, E. and De-La-Hoz-Valdiris. (2022), ‘Trends and Future Perspective Challenges in Big Data’, In: Pan, JS., Balas, V.E., Chen, CM. (eds) *Advances in Intelligent Data Analysis and Applications. Smart Innovation, Systems and Technologies*, (253), Singapore: Springer. https://doi.org/10.1007/978-981-16-5036-9_30

Noy, C., and Hamo, M. (2019), ‘Stance-taking and participation framework in museum commenting platforms: On subjects, objects, authors, and principals’, *Language in Society*, 48(2), 285-308.

O'Halloran, K. ed. (2004), *Multimodal Discourse Analysis: Systemic Functional Perspectives*. London: Continuum.

Poulsen, S. V., & Kvåle, G. (2018), 'Studying social media as semiotic technology: a social semiotic multimodal framework', *Social Semiotics*, 28(5), 700-717. https://doi.org/10.1080/10350330.2018.1505689

Rachunok, B., Fan, C., Lee, R., Nateghi, R., & Mostafavi, A. (2022), 'Is the data suitable? The comparison of keyword versus location filters in crisis informatics using Twitter data', *International Journal of Information Management Data Insights*, 2(1). https://doi.org/10.1016/j.jjimei.2022.100063

Reddy, S.M., Agarwal, S. (2021), 'Isn't It Ironic, Don't You Think?' In: Mantoro, T., Lee, M., Ayu, M.A., Wong, K.W., Hidayanto, A.N. (eds) *Neural Information Processing. ICONIP 2021. Lecture Notes in Computer Science*, 13111. Cham: Springer. https://doi.org/10.1007/978-3-030-92273-3_43

Ruiz-Soler, J. (2017), 'Twitter research for social scientists: A brief introduction to the benefits, limitations and tools for analysing Twitter data', *Revista Digitos* (3) 17-31.

Safa, R., Bayat, P. & Moghtader, L. (2022), 'Automatic detection of depression symptoms in twitter using multimodal analysis', *J Supercomput*, 78: 4709–4744 https://doi.org/10.1007/s11227-021-04040-8

Saif, H., He, Y., Fernandez, M., & Alani, H. (2016), 'Contextual semantics for sentiment analysis of Twitter', *Information Processing & Management*, 52(1): 5-19. https://doi.org/10.1016/j.ipm.2015.01.005

Saniei, R., & Rodríguez Doncel, V. (2022), 'PHDD: Corpus of Physical Health Data Disclosure on Twitter During COVID-19 Pandemic', *SN Computer Science*, 3(3): 1-10. https://doi.org/10.1007/s42979-022-01097-

Sardinha, T. B. (2022), 'Corpus linguistics and the study of social media: a case study using multi-dimensional analysis', In McCarthy. M., and O'Keeffe, A. (eds) *The Routledge Handbook of Corpus Linguistics*, 656-674, London: Routledge.

Saura, J. R., Ribeiro-Soriano, D., & Saldaña, P. Z. (2022), 'Exploring the challenges of remote work on Twitter users' sentiments: From digital technology development to a post-pandemic era', *Journal of Business Research*, 142: 242-254. https://doi.org/10.1016/j.jbusres.2021.12.052

Stockl, H. (2004), 'In between modes: Language and image in printed media', *Perspectives on multimodality*, 1: 9-30.

Tufekci, Z. (2014), 'Big questions for social media big data: Representativeness, validity and other methodological pitfalls', In *Eighth international AAAI conference on weblogs and social media*.

Van Atteveldt, W., van der Velden, M. A., & Boukes, M. (2021), 'The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms', *Communication Methods and Measures,* 15(2): 121-140. https://doi.org/10.1080/19312458.2020.1869198

Van Hee, C., Lefever, E., & Hoste, V. (2018), 'We usually don't like going to the dentist: Using common sense to detect irony on Twitter', *Computational Linguistics*, 44(4): 793-832. https://doi.org/10.1162/colia00337

Weinberg, C., & Gordon, A. S. (2015), 'Insights on privacy and ethics from the web's most prolific storytellers', *The 7th Annual ACM Web Science Conference (WebSci '15).* Oxford, UK: Association for Computing Machinery. http://dx.doi.org/10.1145/2786451.2786474

Weller, K. (2014), 'What do we get from Twitter-and what not? A close look at Twitter research in the social sciences', *Knowledge Organization,* 41(3): 1-15.

Wortham, S., & Reyes, A. (2020), *Discourse analysis beyond the speech event,* London: Routledge.

Yaqub, U., Sharma, N., Pabreja, R., Chun, S. A., Atluri, V., & Vaidya, J. (2018), 'Analysis and visualization of subjectivity and polarity of Twitter location data', In *Proceedings of the 19th annual international conference on digital government research: governance in the data* age, 1-10. https://doi.org/10.1145/3209281.3209313

Zappavigna, M., & Logi, L. (2021), 'Emoji in social media discourse about working from home', *Discourse, Context & Media,* 44, 100543. https://doi.org/10.1016/j.dcm.2021.100543

Zhang, C., Yu, M. C., & Marin, S. (2021), 'Exploring public sentiment on enforced remote work during COVID-19', *Journal of Applied Psychology*, 106(6): 797. https://doi.org/10.1037/apl0000933

Zhang, Y., Chen, F., & Rohe, K. (2022), 'Social Media Public Opinion as Flocks in a Murmuration: Conceptualizing and Measuring Opinion Expression on Social Media', *Journal of Computer-Mediated Communication,* 27(1). https://doi.org/10.1093/jcmc/zmab021

Zimmer, M., & Proferes, N. J. (2014), 'A topology of Twitter research: disciplines, methods, and ethics', *Aslib Journal of Information Management,* 66(3), 250-261. https://doi.org/10.1108/AJIM-09-2013-0083