Applied Corpus Linguistics 5 (2025) 100118

Contents lists available at ScienceDirect

Applied Corpus Linguistics

journal homepage: www.elsevier.com/locate/acorp

Articles

Is LIWC reliable, efficient, and effective for the analysis of large online datasets in forensic and security contexts?

Madison Hunter *, Tim Grant

Aston Institute for Forensic Linguistics, Aston University, Aston Street, Birmingham, B4 7ET, UK

ARTICLE INFO

Keywords:
LIWC
Reliability
Computerized text analysis
Forensic linguistics
Discourse analysis

ABSTRACT

This article evaluates the reliability, efficiency, and effectiveness of Linguistic Inquiry and Word Count (LIWC; Boyd et al., 2022) for the analysis of a white nationalist forum. This is important because LIWC has been the computational tool of choice for scores of studies generally and many examining extremist content in a forensic or security context. Our purpose, therefore, is to understand whether LIWC can be depended upon for large-scale analyses; we initially examine this here using a small sample of posts from a set of just eight users and manually checking the program's automated codings of a subset of categories. Our results show that the LIWC coding cannot be relied upon – precision falls to as low as 49.6 % and recall as low as 41.7 % for some categories. It would be possible to engage in considerable manual correction of these results, but this undermines its purported efficiency for large datasets.

Introduction

Is *Linguistic Inquiry and Word Count* –LIWC (Boyd et al., 2022) –reliable, efficient, and effective for the analysis of large online datasets in forensic and security contexts?

No.

Or at least, LIWC cannot be depended upon for accuracy nor be used without considerable manual analysis, thereby reducing its efficiency. In combination, these two factors reduce the system's effectiveness for the analysis of large online datasets in forensic and security contexts. This is important as there are thousands of Google Scholar hits that indicate LIWC is being used in these contexts, and none that could be found that consider this question.

In forensic and security contexts, an important practice is the linguistic and psychological analysis of extremist texts and interactions that can be found online. An aim of these analyses can be to achieve a deeper understanding of individual and community identities, interests, and ideology through examining the language that those individuals use to express their views and relate to one another. This type of forum represents a particularly important domain to investigate as there has been strong evidence showing that individuals who have committed violent crimes (e.g., Anders Brevik) have expressed their beliefs on such fora. As such research on deviant online communities such as this far right fora, are increasingly seen within the study of forensic linguistics (see e.g. Newsome-Chandler and Grant, 2024) as a contribution to tackling online crime. In addition to linguistic research, approaches can be psychological (e.g., Miron and Douglas, 1979), or sociological, such as social actor analysis (van Leeuwen, 1996) as used by Booth (2023b). Methods in such work can then be manual (Shuy, 2014), computer assisted (Baker and Vessey, 2018), or wholly computational (Pennebaker and Chung, 2009).

Manual approaches –which might include discourse analysis approaches such as appraisal analysis from within linguistics (Martin and White, 2005), or content/thematic analysis from within psychology (Braun and Clarke, 2006; Gottschalk et al., 1958) –have been argued to be more time-consuming, less reliable, and less objective than computational approaches (Tausczik and Pennebaker, 2010). Contrarily, however, it might be argued that they have an advantage over computerized text analysis (CTA) approaches in that they can provide immense detail about the patterns of **function** and meaning –i.e., in how linguistic resources are employed to create meanings, rather than just what linguistic **forms** (words, collocations, syntactic structures) are most frequent (Biber et al., 1998; Gales, 2010; Tayebi, 2016).

Computational text analyses –including LIWC –are preferred by their advocates because they provide a fast and apparently reliable means to examine large quantities of text to extract semantic information just based on counts of word forms (Tausczik and Pennebaker, 2010). The key advantage is that methods are replicable and this can be an

* Corresponding author at: Aston University, Aston Street, Birmingham, B4 7ET, UK.
E-mail address: m.hunter5@aston.ac.uk (M. Hunter).

https://doi.org/10.1016/j.acorp.2025.100118
Received 17 July 2024; Received in revised form 22 December 2024; Accepted 8 January 2025
Available online 9 January 2025

important criteria for some research and operational applications in forensic and security contexts. There can be a tension however between this reliability and the validity of methods (Grant and Baker, 2001). CTA approaches, have been critiqued as not being able to provide the kind of "nuanced understanding" (Gales, 2010, p. 53) of functional language patterns that manual approaches can offer. Within psychology, the wholly automated LIWC software (Boyd et al., 2022) has become the increasingly popular choice for a CTA approach. Part of LIWC's appeal is what it was designed to do: identify words or short phrases which represent specific psychological constructs (e.g., Pennebaker et al., 2014). In other words, unlike discourse analytic approaches, LIWC is meant to automatically analyze form **and** function at the same time and –in some cases explicitly, in others implicitly –to give a window into the psychological drivers behind usage. It does so by classifying each word or phrase into one or more of its many "psychologically meaningful categories" (Tausczik and Pennebaker, 2010, p. 24), then producing an output that shows the percentage of total words in a text that fall under each category.

Being a popular choice for CTA, LIWC is very widely used across a range of research contexts, not just the psychological context in which it was created. The creators themselves cited that a broad Google Scholar search for 'LIWC text analysis' yielded over 2400 results from **just 2021** (Boyd et al., 2022, p. 13). Replicating this approach for estimating the significance of LIWC research confirmed that Google Scholar produced 27,000 hits in total and running more focused searches with queries such as 'LIWC forensics' or 'LIWC security', produces roughly 1500 and 7500 hits, respectively; and an even more specific query of 'LIWC extremism' points to nearly 1000 references to academic papers.

Despite the claims about, and apparent demonstrations of, LIWC's reliability, efficiency, and effectiveness as a method for learning more about an author from their language (e.g., Chung and Pennebaker, 2018b; Tausczik and Pennebaker, 2010), direct evaluations of these claims seem almost exclusively to have been done by the originators and proponents of the system. Even these evaluations are narrow in focus examining the original development within the original research context, even though subsequent application of the software has been very broad. There are some exceptions in the linguistics literature; Lorenzo-Dus and Kinzel (2019) reject the use of LIWC because of "non-transparency of some of its analytic variables and a focus on de-contextualized words" and Gillings (2024) has similar concerns, both from a corpus assisted discourse analysis perspective. It remains the case, however, that an independent systematic evaluation of the reliability of LIWC has not been carried out. With these observations in mind, this paper describes the application of LIWC to interactions on a white nationalist, extremist, online forum to determine the benefits and limitations of the method, when applied to data relevant to forensic and security contexts.

LIWC

Originally developed in the 1990s as part of research into the relationship between physical and psychological health and language use (Francis and Pennebaker, 1992), LIWC is currently on its fifth iteration (LIWC-22; Boyd et al., 2022). In every iteration of LIWC, there has been, at least, a main text processing module, which works by moving sequentially word-by-word through each text in a dataset and assigning each word or short phrase to one or more of the over 100 categories. This is determined by matching that word or phrase to an entry in LIWC's internal dictionary of "over 12,000 words, word stems, phrases, and select emoticons" (Boyd et al., 2022, p. 5), which have all been tagged with the categories relevant to them. Before this can be done, though, the texts must be 'cleaned' to remove spelling errors (including misspellings and nonstandard spellings) and other formatting problems (e.g., it cannot read words separated by punctuation with no space like *dog/cat*) .[1] The main analysis output then shows the percentage of the total words in the text that fall into each category. The exceptions to this are the 'summary variables' –*clout, authenticity, analytical thinking,* and *emotional tone* –which use proprietary algorithms that combine various LIWC categories to calculate percentile scores (Boyd et al., 2022). In addition to the traditional LIWC analysis module, eight 'companion modules' were added to LIWC-22 to allow users to analyze their texts in different ways. However, for the purposes of this paper, the focus is placed on the traditional LIWC analysis module as it has been the central feature of the software for all its iterations.

LIWC's reach into research on topics relevant in forensic and security contexts has become extensive, its application in these domains is of significant scale, and the questions addressed are potentially very important. Much of the research has an applied focus with an eye toward providing methods of analysis that would be useful to law enforcement and security services for investigative and preventative purposes (Chung and Pennebaker, 2018a). Applications include examination of far-right extremism and radicalization (Figea et al., 2016; Grover and Mark, 2019) and related topics like ideology-driven violence (Kaati et al., 2016), terrorism (Pennebaker and Chung, 2009), and risk assessment (Glasgow and Schouten, 2014). Analysis of far-right extremist ideology has become increasingly critical in recent years as the support for the ideology has increased (Mudde, 2019) and its reach has expanded to a global audience with the rise of anonymous web fora (Baumgarten, 2017). Currently, far-right extremism presents a growing threat worldwide, with terrorist attacks committed by subscribers to the ideology accounting for 30 percent of all terrorism-related fatalities in the West between 2007 and 2021 (Institute for Economics and Peace, 2022).

In application to extremist texts generally, including analyses of the far-right, there are a concentrated group of LIWC categories that receive the most attention. These are:

- ***anger*** –words relating to emotional and behavioral expressions of anger;
- ***emotion*** –words describing feelings and behaviors associated with common emotions;
- ***power*** –words depicting power dynamics;
- ***third-person plural pronouns*** (sometimes ***pronouns*** generally); and
- ***big words*** –those with six or more letters.

Anger and *negative emotion words* more broadly have been found to be higher for individuals with radical/extremist ideologies and those who commit acts of violence (Grover and Mark, 2019; Pennebaker and Chung, 2009). The argument is that they are common because these individuals harbor intense dissatisfaction with some larger entity (e.g., society, government, a particular race). High frequencies of *power* words are argued to indicate a need for power and authority over others (e.g., Kaati et al., 2016). It is argued that a higher frequency of *third-person plural pronouns* may indicate a greater amount of 'othering' and reflect an 'us versus them' mindset (e.g., Pennebaker and Chung, 2009). Finally, *big words* are argued to indicate cognitive complexity and have been found to be negatively correlated with emotionality (Baele, 2017; Kaati et al., 2016). Many of these studies contend that, by using LIWC to identify patterns of psychologically meaningful linguistic features, we

[1] From a linguistic research perspective and a forensic linguistic evidential perspective, such data cleaning (as is common in computational approaches) might be considered problematic in itself: such practices necessarily alter the data, making it no longer truly natural language, and could indeed be considered contamination or alteration of evidence if ever used in legal disputes.

can begin to deepen our understanding of the authors and the groups to which they belong (Glasgow and Schouten, 2014; Pennebaker and Chung, 2009).

Three immediate observations can be made when reviewing this body of research, which in combination suggest a lack of critical examination of LIWC, and which motivated the investigation that led to this paper. The first observation is that many studies seem to cite LIWC's widespread use in research as part of the justification for employing it (e.g., Drouin et al., 2017; Glasgow and Schouten, 2014; Torregrosa et al., 2020).

Relatedly, the second observation is that, despite the clear widespread use, there appears to be a lot of circular citations within the literature. That is, many LIWC studies cite other LIWC studies as providing the basis for investigating links between language use and psychological characteristics (e.g., Kacewicz et al., 2014; Pennebaker et al., 2014). These practices are not necessarily problematic in themselves more generally, but in this case, they are because they occur despite seemingly minimal independent evaluation of the program's functionality being carried out, apart from the reliability and validity tests run by the creators themselves (e.g., Boyd et al., 2022) and the 'field-testing' done in the early stages of development (e.g., Pennebaker and King, 1999). It should be noted that the earlier iterations (i.e., the 2001 and 2007 versions) appear to have essentially been 'black boxes' (which is itself problematic) in that there was not a way to examine context or what the program coded under each category, thus making it impossible to independently evaluate performance. However, there has been at least a color-coding function to examine context with since the 2015 version (Pennebaker et al., 2015), meaning that the means for independent assessment has existed for nearly a decade, but researchers have seemingly not capitalized on it.

The final observation is that it appears that the suitability of LIWC for analyses on text types far removed from those that were used to create its dictionary (like far-right extremist forum posts) is only briefly acknowledged in some of these studies (e.g., Figea et al., 2016; Glasgow and Schouten, 2014) but not thoroughly investigated. It was built using texts from sources such as NYT articles, Twitter posts, and books from Project Gutenberg (Boyd et al., 2022) that are subject to more scrutiny and greater restrictions on content than extremist fora, and which differ from such content in at least one or more other factors like audience, mode of communication, and genre. All these contextual factors can have a great deal of influence on the meaning-making resources with which authors choose to express themselves (Bell, 1984; Biber et al., 1999; Halliday and Matthiessen, 2014), which would likely impact the LIWC output and, in turn, the conclusions that can be drawn from it.

These three observations suggest a lack of empirical rigor that, at the very least, presents a major methodological concern, and at most, poses a not-so-insignificant risk given the potential high-stakes applications of the research within forensic and security contexts.

There are also limitations with the main LIWC analysis module itself that may raise concern from a linguistic analysis perspective. The biggest limitations of the module stem from how it annotates the data, which as mentioned above, is done wholly automatically by working sequentially through the text word-by-word (and some short phrases in LIWC-22; Boyd et al., 2022). The result is a speedier analysis, but the cost is that context and its impact on meaning are not considered. Context plays a crucial role in making and determining meaning (Halliday and Matthiessen, 2014), and the consequences to being unable to consider its influence are (at least) threefold. The first two relate to what is lost when analyses are concerned with just decontextualized linguistic forms; namely, that there is (1) an inability to fully capture implicit meanings and (2) a higher risk of misclassifying forms which have more than one possible meaning. Regarding the first point, not all meanings are conveyed explicitly; an author can use shared knowledge and contextual cues to imply a particular feeling or judgment with language that is inherently neutral, ambiguous, or non-evaluative (Bednarek, 2009; Don, 2016; Tayebi, 2016). This may include, for example, comparing someone to a movie character to imply a positive or negative judgment or the use of group-specific language (Don, 2016), such as how white nationalist groups use numbers to represent slogans conveying hateful or derogatory beliefs (Anti-Defamation League, 2023). Such context-dependent meanings would necessarily be missed by the LIWC analysis as the value assigned to the words or phrases relies on shared knowledge.

Regarding the second point, context helps resolve ambiguity between words that look the same (i.e., same form), but may take on different meanings (i.e., functions) in different environments (Biber et al., 1998). This limitation has been acknowledged by the LIWC creators (e.g., Tausczik and Pennebaker, 2010), but they argue the word-count strategy assumes that words carry "psychological information over and above their literal meaning and independent of their semantic context" (Chung and Pennebaker, 2011, p. 7). They also defend the program by saying that LIWC uses probabilistic models of language use, meaning that with more data, the impact of misclassifications would be lessened (Pennebaker Conglomerates, n.d.), using the word *mad* as an illustrative example. *Mad* is included in the 'anger' category in LIWC, but it can also denote enthusiasm (*they're mad about each other*), serve as an intensifier (*I'm mad happy*), or describe someone as mentally unstable (*he went mad*). The probabilistic models of language use would suggest if the positive meaning of *mad* is meant, it is likely that other positive terms will be used elsewhere in the text and that the misclassification would not have a significant impact on the overall distributions observed for that data. However, this would not necessarily always be accurate, as different communicative contexts will impact the choices language users see as available to them in different ways based on factors like audience (Bell, 1984), genre (Swales, 1990), or topic (Biber et al., 1999). A further aspect of this which has received less attention is the resolution of syntactic ambiguity –verbs such as 'to go' and 'to have' for example, can function either as a main verb or an auxiliary verb and will carry very different meanings and functions accordingly. Discourse markers such as 'so' have many varied uses such as intensifier (so great), prompt for another person to speak (So?), summarizer, or new topic introducer etc. Failure to factor in discoursal function misses potentially significant linguistic evidence, including those useful for assessing personality. As LIWC does not tag the text for part of speech or attempt to parse it (standard approaches early in many CTA pipelines), both easy-to-resolve and harder ambiguities remain unresolved.

Finally, because the LIWC categories are all separate and the program cannot consider contextual influence, the output cannot provide information about the relationships between elements in the texts. This is particularly problematic for being able to identify whether something was negated and the source and target of any evaluative language. All of these are details that are important for interpreting the linguistic patterns (Du Bois, 2007) and without them, the account of the data is necessarily incomplete. Knowing the source and target of evaluations would allow one to determine what proportion of the evaluation-related categories capture the author's personal views (versus those they attribute to, or claim to share with, others) and to whom/what the views relate. Regarding negation, multiple problems arise including that the output cannot show what types of meanings are most often negated, it is unable to capture the difference between some non-negated (e.g., *happy*) and negated (e.g., *not happy*) expressions, and it cannot identify instances of polarity reversal from negation earlier in the sentence (e.g., *I don't think I am upset*).

Without the above pieces of contextual information or a subsequent qualitative analysis, it becomes more difficult to determine the importance of any parts of the LIWC output, interpret the findings, and generate any conclusions based off them. This is particularly true with texts that have contextual factors that differ greatly from those used to create the LIWC dictionary, such as online far-right extremist fora posts.

Data and methods

Dataset

The corpus from which the data were selected was compiled by members of the Aston Institute for Forensic Linguistics and comprises posts amounting to 417 million words from around 113,000 usernames[2] of a white nationalist[3] forum spanning the years 2001–2018. As the intent of this study was to check LIWC coding manually, criteria were developed to help select a sample small enough to be manually analyzed within a reasonable time and resource frame. In other words, the goal here is a rich data analysis, not so much a big data analysis, with a view to finding examples to tests where LIWC was or was not operating as claimed. The choice to analyze just a small sample created some risk: if no real problems were found, it would not have been able to provide strong positive evidence that LIWC worked well against the claims made for it, but if problems were found with a small sample, this could reliably be used to infer a more widespread problem. Our approach rests on the principle of falsification –a single counter-example, or small number of counter-examples may be enough to suggest that an approach or theory is flawed, or requires further development before being used.

The sample texts used in this analysis thus came from just **eight** users who collectively posted at different points over a period of ten years of the forum's existence. Following Booth (2023a) work on 'careers' (essentially posting duration/frequency) within such fora, we selected two users who posted within just one year; two who posted within a two-year period; two who posted over a three-to-six-year period; and two who posted for seven or more years. Within each of the four career types, users were selected using a random number generator and sufficient posts for each user were randomly selected to amount to a minimum of 250 words. While it is not a large dataset, this sample did allow for a more in-depth exploration of how LIWC performs when analyzing this genre of interaction. The selection of users with different posting durations is intended to simply control for any linguistic differences that may occur as longer-term participants in their fora might acculturate and change their linguistic practices (Booth, 2023a) in such a way as could affect LIWC codings. The details of the sample are included in Table 1.

Table 1
Breakdown of dataset.

User #	Career length	Selected Posts Years	Word count
1	2009–2018 (10 years)	2009 / 2012 / 2015 / 2018	323
2	2009–2018 (10 years)	2009 / 2012 / 2015 / 2018	366
			689
3	2009–2011 (3 years)	2009 / 2010 / 2011	405
4	2014–2016 (3 years)	2014 / 2015 / 2016	380
			785
5	2009–2010 (2 years)	2009 / 2010	307
6	2014–2015 (2 years)	2014 / 2015	266
			573
7	2009 (1 year)	2009	343
8	2015 (1 year)	2015	381
			724
		Total	**2771**

[2] Although it is recognized that one individual may have multiple usernames and one username might be used by multiple individuals, this is an unimportant distinction for the current study and for convenience we use 'user' throughout to indicate username.

[3] 'White nationalist' is a self-identifying label for a sub-set of far-right extremists (Booth, 2023b) and we follow the convention of using the self-identifier to refer to the group collectively, without commitment to or comment on its potentially prejudicial connotations.

Methods

The newest version of LIWC was used (LIWC-22; Boyd et al., 2022) for this analysis, as it is the most up-to-date version available. As stipulated in the program's manual, spelling errors (misspellings and nonstandard spellings) and other formatting problems (e.g., words separated by punctuation without spaces like *dog/cat*) were fixed in all texts prior to uploading them to the program. This is to ensure that it reads and classifies as many words as possible.

As mentioned in the above section, several companion modules were added to LIWC-22 (Boyd et al., 2022). The main analysis module alone does not provide a way to examine the results qualitatively, which as will be seen, was a necessary aspect of assessing LIWC's performance. Therefore, although the focus of this analysis is to evaluate the central LIWC module, two of the companion modules –the 'case study' and 'contextualizer' –were used to aid in this. The case study module allows individual files to be processed and has a color-coding option that highlights all words/phrases coded under a particular category in the text. The color-coding option is not completely new as it was available in the 2015 version of LIWC (Pennebaker et al., 2015), but in a seemingly less sophisticated form and not as part of a larger module. The contextualizer displays all the words/phrases coded for a chosen category within its immediate context and provides a count of how many instances of a given category there are in a text (Boyd et al., 2022). Additionally, the creators provide a full downloadable list of all words in each category within the program; this was used to help properly track the error types described below.

For each of the eight authors, all posts were placed into a single document. The individual posts were labelled with single-word tags at the top indicating the author and year so they could be distinguished in the qualitative analysis. As there are over 100 total categories in the program, the focus needed to be narrowed to a more reasonably-sized set. Two groups of LIWC categories were selected as sample codes for analysis –the most frequently used categories and the most important categories that arose out of the literature that applied LIWC to extremist texts. Again, the choice to use just a reduced sample of codes creates a greater risk of underestimating problems found in LIWC coding rather than overestimating any such problems. The first step was to identify those categories that were used over 5 % of the time by every author –this yielded a list of 19. Eleven of those were then removed for several reasons: four were summary categories (e.g., *word count*); three were summary variables (e.g., *clout*) which cannot be examined qualitatively as their formulae, and how the base categories are used and weighted within them, are unpublished; three were too broad (e.g., the *Linguistic* category which includes all categories relating to linguistic function); and one, *cognitive processes*, had very similar values to its parent category, *Cognition*, which was already in the list. It might be argued that excluding subcategories in favor of the parent categories is a flawed approach, but it allows a wider examination of LIWC's performance, offering more opportunities for the program to succeed. This left **eight** categories based on the LIWC analysis of this dataset. In addition, the literature review of LIWC papers investigating extremists' writings gave rise to a further **seven** categories, which were identified as important in this topic domain and genre. Thus, a total of **fifteen** categories were examined in detail. These categories can be considered broadly representative of the categorization system of LIWC (three of the categories encompass at least nine other sub-categories) although of course it is not possible to generalize specifically from these categories into others not examined here. The final list is provided in Table 2.

It should be noted that some of the remaining categories in Table 2 are subsets of other categories in the list. For the frequency-based group, *personal pronouns* and *auxiliary verbs* are subsets of *pronouns* and *verbs,* respectively; for the literature-based group, *positive emotion, negative emotion,* and *anger* are subsets of *emotion* (which is itself a subset of *affect*). The frequency-based group subcategories were kept because they feed into at least one of the summary variables (Pennebaker et al.,

Table 2
List of analyzed LIWC categories.

	Category	Description
Frequency-based selection	**Pronoun**	All pronouns including personal (*I, you, we*) and impersonal (*this, that*)
	Personal Pronouns	Pronouns that refer to a person (e.g., *I, you, we*)
	Determiners	Words that precede nouns to specify or clarify its meaning/scope
	Prepositions	e.g., *to, from, above, across*
	Auxiliary Verbs	'Helping' verbs that precede main verbs and provide grammatical information like tense or modality (e.g., *have, be, do, can, might*)
	Verbs	Includes a set list of 'common' main verbs
	Cognition	Words for describing/referring to thinking processes
	Social	Words for describing socially relevant behaviors, roles, and relationships
Literature-based selection	**Affect**	Words for expressing positive or negative feelings or judgments
	Emotion	Words for describing emotions (includes positive and negative)
	Negative Emotion	Words for describing negative emotions
	Positive Emotion	Words for describing positive emotions
	Anger	Words relating to the feeling and expression of anger
	Power	Words for describing power dynamics
	3rd Person Plural	i.e., *they/them* and related forms

2014), meaning that frequent errors could adversely impact their reliability. The literature-based group subcategories were kept primarily because they are so common in the literature. However, it was also because, unlike the *cognitive processes* category, there is not as much overlap with their superordinate category (e.g., **affect** includes emotions, but also more general positively- and negatively-valenced words), so analyzing them separately allows examination of the program's performance at broader and narrower levels of categorization of evaluative language. It should be noted that instances (tokens) counted as errors in lower-level categories **were not** counted more than once for their superordinate categories—that is, if an error occurred on the same word within *anger* and *emotion,* it was only counted once for the error total for *affect*, not three times. Using the detailed record of categorizations, these counts were double-checked before the accuracy statistics were calculated to make sure that the same error was not counted more than once for these hierarchically-related categories.

Once the 15 categories were identified, the documents for each author were uploaded one at a time for analysis. Since the LIWC analysis module output only provides percentages, the contextualizer was used to find out the exact number of tokens that the program coded under each category. Each category was selected one-by-one with the color-coding function so the tokens could be examined more closely within their context and notes were taken about any observed errors to keep track of their frequency.

Categorizing errors

Two broad types of errors were noted (one was realized in two different ways). They are introduced here for reference but are discussed in more detail in the results section.

False Negatives: This error type occurred when a word/phrase that should have been coded by LIWC was not tagged. This type had two realizations:

a. ***Words not in the relevant LIWC dictionary,* but which had a similar meaning to (and/or represented the same construct as) one or more other entries in that dictionary category.** These were determined using the uncoded words' dictionary definitions and synonyms listed in the *Oxford Languages* dictionary and thesaurus. When a word that appeared to fit within the category being examined was identified in the manual analysis, the LIWC dictionary was checked to see if it was in the category or any related categories. If not, its dictionary definition and synonyms were checked to see if they matched any other members of the relevant category. The decisions are based in partially subjective judgment, but they were initially recorded and then developed into a 'policy book' through discussion with the second author. Additionally, words were not counted as an error if there was not a clear and direct connection to other items in that category either through its dictionary definition or synonyms.

An example of this sort of false negative would be *fight,* which is similar to other words (e.g., *confrontation* or *battle*) in the **power** category but is not included in it and therefore not coded by LIWC. The process ultimately amounted to a search through uncoded items in the text, which would credibly be expected to fall into the coding category under examination based on their meaning and function in context. By checking decisions with a second judge, we tried to be as complete as possible, but are aware that another coding team might allocate additional words into the categories. That is to say, our judgments of this type of false negative are more likely to be an underestimate of actual occurrence than an overestimate.

a. ***Words that were in the relevant LIWC dictionary and therefore should have been coded by LIWC but were not coded***: These occurred when, through a technical coding error (such as a misspelling within the software itself), the program did **not** code a word/phrase that was included in the published LIWC dictionary for the category being examined. This was rare but it did occur. An example would be *weren't* not being coded under the **verb** category despite being included in it. This error type and possible explanations are discussed in more detail in the next section.

False Positive: This type occurred when a word/phrase was coded under a particular category, but it is not being used in that way (e.g., when *that* is tagged as a pronoun but is functioning as a complementizer); in other words, when the program misclassified the word/phrase.

The *false positive* frequencies were subtracted from the overall frequency for each category taken from the contextualizer output, which yielded the total number of **'true positives'** (i.e., the total number LIWC correctly coded). These numbers were then used to generate the **precision** and **recall** statistics which are commonly used as a measure of coding performance (e.g., Read and Carroll, 2012). The precision score shows the percentage of items the program assigned to a particular category that were correctly assigned to that category. The recall score shows the percentage of items that should have been assigned to a category that were correctly assigned. These metrics help calculate the **F-score**, which is an overall measure of the program's accuracy (Chinchor and Sundheim, 1993). The results from these tests are presented in the next section.

Results

The discussion of results requires quotation from the posts which raises issues of both anonymization and the potential to cause offence through the quoting of hate speech. This has been avoided by deliberately truncating quotes and making minor alterations to some examples but only to words surrounding those highlighted in the analysis, which thus would not have altered the LIWC coding.

Table 3 contains the frequencies of total coded items, true positives, false positives, and false negatives for all eight authors combined for

Table 3
Total frequencies for all categories.

LIWC Category	Total coded	True positive (correct)	False positive (misclassified)	False Negative	
				In LIWC	Not in LIWC
Pronouns*	385	325	60	0	10
Personal Pronouns*	231	230	1	0	0
Determiners*	390	311	79	0	1
Prepositions*	352	284	68	0	0
Aux Verbs*	280	139	141	0	1
Verbs*	493	467	25	1	133
Cognition	406	399	6	1	51
Social	400	398	0	2	161
Affect	188	160	6	21	102
Emotion	55	54	1	0	38
Negative Emotion	32	31	1	0	22
Positive Emotion	22	22	0	0	14
Anger	10	10	0	0	14
Power	66	61	3	1	72
3rd Person Plural*	87	87	0	0	0

each LIWC category analyzed. The starred* items are linguistic categories, the unstarred items are psychological process categories, and the large space between the **social** and **affect** categories separates the frequency-derived from the literature-derived categories.

As Table 3 shows, when errors occurred within the linguistic categories, they were mostly of the 'false positive' type (except for **verbs**). When errors occurred within the psychological processes categories, they were mostly 'false negatives', specifically of the sub-type covering instances of words/phrases not in LIWC's dictionary, but which had a similar function to one or more entries in for that category and/or represented the same construct as the category.

The formulae used to calculate the precision, recall, and F-scores (Chinchor and Sundheim, 1993; Read and Carroll, 2012) are as follows:

$$\textbf{\textit{Precision}} = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$\textbf{\textit{Recall}} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$\textbf{\textit{F}} - \textbf{\textit{score}} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Table 4 shows how the error rates in Table 3 translated to these three measures of accuracy. Categories for which precision or recall rates were below 75 % (0.750) are indicated with a superscript 'a'. If precision is lower than 75 %, it means that 75 % or less of the tokens the program marked as belonging to a particular category were properly classified (i.e., a higher false positive rate). If recall is below 75 %, it means that the tokens the program classified into a particular category only represent 75 % or less of the tokens that **should have been** classified into it (i.e., a higher false negative rate). This is an admittedly arbitrary threshold and indeed other measures for accuracy of a system are available in computational research (Vickers et al., 2024). Picking a threshold will always be arbitrary which is why all figures are provided enabling evaluation of LIWC's performance overall but using a threshold to select texts for further examination was a pragmatic solution. The selection of 0.75 also roughly aligns with social science practice that reliability (Cohen's Kappa) of about that level is considered 'substantial' but not perfect agreement (e.g. Landis and Koch (1977) and so provides a useful rule of thumb.

Table 4
Precision, recall, and F-score.

LIWC Category	Precision	Recall	F-Score
Pronoun	0.844	0.970	0.903
Personal Pronouns	0.996	1.000	0.998
Determiners	0.797	0.997	0.886
Prepositions	0.807	1.000	0.893
Aux Verbs	0.496[a]	0.993	0.662
Verbs	0.949	0.777	0.855
Cognition	0.985	0.885	0.932
Social	1.000	0.709[a]	0.830
Affect	0.964	0.608*[a]	0.746*
Emotion	0.982	0.587[a]	0.735
Negative Emotion	0.969	0.585[a]	0.729
Positive Emotion	1.000	0.611[a]	0.759
Anger	1.000	0.417[a]	0.588
Power	0.953	0.455[a]	0.616
3rd Person Plural	1.000	1.000	1.000

The asterisk next to the recall score for **Affect** is because all but one of the 'technical coding' false negatives for that category resulted from a spelling error in the dictionary itself—that is, the lemma for the derogatory term *negro* was misspelled *neegro*. This meant that no instance of the term in the dataset would be properly captured unless the spelling was incorrect in the same way in the texts themselves. As such, the higher 'technical coding' false negative rate for that category represents an outlier that, while important to note, would artificially lower the recall and F-score.

Table 4 demonstrates that the precision rates are lower than the recall rates for many of the linguistic categories –where false positives were more common –and the recall rates are lower than precision rates for many psychological process categories –where false negatives were more common.

For the linguistic categories, the false positives (misclassifications) occurred when the word/phrase that was coded **can** belong in the given category but **does not** in the context in which it appears. That is, the program correctly identified the form but not the grammatical usage. For example:

1	We can merely hope **that** it will not go unnoticed	**(pronoun; user 7)**
2	Even if **some** maybe don't believe every detail	**(determiner; user 4)**
3	They are able to use DNA **to** follow our every move	**(preposition; user 2)**
4	It should **be** about **having** a rebellious attitude	**(aux verb; user 3)**
5	You have to love his **stumbling** and stuttering to smooth it over	**(verb; user 7)**

Each of the above examples represents common realizations of the false positive error for their given category. With **pronouns**, many of the false positives were from forms like *that, this, those,* and *these* (which can function as pronouns) being coded when they were being used for a different function (e.g., as determiners as in *this cake*). In example 1, *that* serves as a complementizer –i.e., introduces a complement clause (Biber et al., 1999) –not a pronoun. Similarly, some forms that can serve independently as pronouns can also be determiners –e.g., *this, that, some* –which was the common realization of false positives for the **determiner** category, as in example 2. With **prepositions**, many of the false positives occurred with the infinitival *to*, as in example 3, and when prepositional forms served as part of a fixed, non-prepositional phrase (and therefore were not functioning as prepositions), like *of course*. For **auxiliary verbs**, all but one of the false positives was due to the program coding the form as an auxiliary when it was functioning as a main verb, as in example 4. Finally, many of the **verb** misclassifications were the result of coding a form that was functioning as another part of speech, such as a noun (example 5) or an adjective (e.g., *the coming months*).

In the psychological process categories (and the **verb** category), there were fewer false positives, but more false negatives, lowering the recall rate. There were three broad patterns of false negatives: (1) words/phrases that had functions related to the category and/or other entries in it; (2) words/phrases not included despite being variations or

derivations of forms that **are** in the dictionary; and (3) words/phrases not captured due to inflexibility within the program.

The first pattern is demonstrated in the following examples:

6	**Work** nonstop to **dispossess**…**ridicule**…	**(verb; user 2)**
7	The disease is **imposed** upon them	**(cognition; user 8)**
8	No **media** nor **propaganda** that is able to **trick** you	**(social; user 6)**
9	Nothing else **matters**	**(affect; user 4)**
10	He ultimately loses the **fight** badly	**(power; user 5)**

With the **verb** category, many of the tokens belonging to the first pattern, like example 6, resulted from the program using a finite list of verbs instead of a part-of-speech tagger. As verbs are an open class (i.e., a word class that allows 'indefinite and unlimited' membership; Biber et al., 1999, p. 56), the use of a finite list increases the risk of missed tokens more than it would with closed-class categories (i.e., those with limited membership) like prepositions.

For the remaining categories, most of the tokens in the first pattern were words/phrases which were semantically similar to other entries in the category. In example 7, the word *imposed* was not classified under **cognition** despite similar words/synonyms like *force* and *inflict* being included. Other examples include words relating to truth values (e.g., *fact* is not included, but *true* and *false* are) or various cognitive processes (e.g., *waking up* and *agenda* are not included, but *aware* and *plan* are).

Example 8 shows three instances within one sentence that were not classified under the **social** category despite denoting behaviors/phenomena which directly relate to aspects of social interactions and relationships. That is, *media* and *propaganda* relate to the dissemination of information to members of a society, and *trick* is semantically related to *fool*, which is in the dictionary, **and** it represents a socially relevant behavior. The **social** category also includes several violent actions and judgments of morality/ethics (e.g., *kill, attack, depraved*), but excludes several others (e.g., *impoverish, genocide, degenerate*).

In example 9, *matters* was not coded under **affect** despite words with similar meanings, like *important*, being included in the category. As this category includes positive and negative judgments, there was a lot of overlap between the false negatives for it and other categories, such as the violent actions and judgments of morality that were an issue with the **social** category. Other false negatives include various derogatory slurs used to reference out-groups (in addition to the one caused by a spelling error discussed above); expressions of desire/intent (e.g., *want/wish* are not included but similar meanings like *would like* are); and expressions of remorse (e.g., *sorry* is not included, but *regret* and *guilt* are).

Example 10 shows one instance of a word that has a similar meaning to other forms in the **power** category –e.g., *confrontation* and *battle* –not being captured. Other acts of incapacitation –which necessarily encode a situation of someone's power being removed (Hurt, 2020) –were also not included in the dictionary, (e.g., *exterminate, beat up, bondage*), despite other acts like *kill* and *genocide* being included. In addition to this, some words that convey power being given to someone are missed, such as *help* and *free* (as in *free a person*); and despite *apologize* being included, *sorry* is not captured.

The second pattern of false negative errors –uncaptured variations/derivations of included words –occurred at least five times in total in each of the **verb, cognition, social, affect,** and **power** categories. For example:

11	Never tried to **teach** them	**(verb; user 8)**
12	They **don't** think they do	**(cognition; user 1)**
13	They don't care who's **right** or wrong	**(social; user 5)**
14	Their **opponent's** moves	**(affect; user 6)**
15	That is why he **lost**	**(power; user 4)**

In each of the above examples, words which have a similar or identical root to the bolded and underlined word are included in the dictionary for the given category. Within the **verb** category, other conjugations of verbs like *teach* were included –e.g., *teaches, taught* –just not the one used in the text. For **cognition,** instances like example 12 were common because most 'do + not' constructions (except *didn't*) were paired with something else (e.g., *do not have*). Other examples included *selection* and *claim/claims* (verb forms) not being captured despite *select* (and its conjugations) and *claimed* being included in the dictionary. With the **social** category, instances like example 13 occurred (*rightly* is included, *right* is not), though false negatives more commonly occurred when racial groups were described with a noun and not an adjective (e.g., *whites*) despite phrases like *white person/people* being included in the category. Within **affect**, examples included instances like 14, where a variation of the term was captured (in this case, *opposition*) but not the one used in the text. Finally, within **power**, there were instances, like example 15, where the nominalized form (*loser*) was included, but not the verb form. Since other verbs are included in the category (e.g., *assault, forbid*), its exclusion is unlikely to be wholly due to its part of speech.

The third and final pattern, which was the least common, were the 'technical coding' false negatives that resulted from inflexibility within the program. For example:

16	I know what **you** mean	**(social; user 1)**
17	A civil **war** or worse	**(affect; user 2)**
18	**Weren't** all bastards	**(verb; user 8)**

The bolded and underlined words in each of the above examples are the words that should have been coded by LIWC (they are all included in the category indicated to the right) but were not, likely because of the program's newly added capability to code certain short phrases. Based on the observations of this error type, it seems that if a phrase is coded as a unit for one category, none of the phrase's component parts can be coded separately under any other categories. Thus, *you* in (16) cannot be coded under **social** because *you mean* is classified under **cognition;** *war* in (17) cannot be classified under **affect** because *civil war* is classified under **power** (despite *civil war* being a type of war); and *weren't* in (18) is not coded as a **verb** because **n't all* is a lemma classified under **cognition**.

Discussion

The biggest of LIWC's limitations stem from its wholly automated word-by-word (and occasional short phrase; Boyd et al., 2022) annotation approach which necessarily ignores context and its influence on meaning. The resulting consequences are at least threefold: (1) implicit meanings are missed, (2) there is a higher risk of misclassification of forms with multiple functions, and (3) information about the relationship between linguistic elements is lost. These consequences arguably become more problematic when analyzing texts that differ across multiple contextual factors to the texts that were used to create the LIWC dictionary. The creators' position is (1) that the program's use of probabilistic language models means that greater amounts of data will likely reduce the impact of misclassifications from not considering context (Pennebaker Conglomerates, n.d.) and (2) that words carry 'psychological information over and above their literal meaning and independent of their semantic context' (Chung and Pennebaker, 2011, p. 7).

This position severely downplays the crucial role that context plays in making and determining meaning (Halliday and Matthiessen, 2014) and, as Bakhtin (1986) argued, "words belong to nobody, and in themselves they evaluate nothing" (p. 85); the evaluative and expressive aspect of words is gained through their use by the speaker in context. Moreover, more thorough evaluations of the program's performance do not appear to have been conducted to corroborate the defense despite its widespread use in research across myriad contexts.

Thus, the purpose of this paper was to explore the performance of LIWC's main analysis module more thoroughly, specifically with data that differs so drastically from that which helped form the dictionary.

This was done by qualitatively examining 15 LIWC categories across a dataset of white nationalist forum posts. Despite the small dataset and range of categories, the qualitative analysis reveals observable patterns of errors that were consistent across categories **and** authors. As the goal of this paper, as stated in earlier, was a rich data analysis, not a big data analysis, we do acknowledge the limitations presented by the small dataset size in terms of our ability to generalize the findings. However, as we have argued previously, and argue further here, that these errors are consistently and frequently found in such a small sample of both text and authors indicates that there may well be a broader problem with the reliability of coding. This should, of course, still be more fully evaluated for different genres and different sets of LIWC categories but as it stands we believe that our analysis is sufficient to raise serious questions as to the reliability of LIWC coding such that it should not be used without the reliability of coding being established through some process of checking that its counts are accurate and valid for the purpose that it is being used.

Amongst the linguistic categories, false positives were relatively low for **personal pronouns** and **3rd person plural pronouns,** and thus did not contribute to the observed pattern of errors. This is unsurprising given that they are narrower and less varied categories than the others considered here. The **3rd person plural pronouns** specifically have been argued to indicate a greater amount of "othering" and reflect an "us versus them" mindset in the context of extremism and terrorism (Pennebaker and Chung, 2009). In this dataset, this category accounted for between 1 % and 5 % of total words for each author. Although the false positives are low and LIWC is doing its job in this respect, consideration of the functional level of understanding of othering may indicate that the recall is also low, which is a problem for using LIWC in this application to extremist discussion fora. That is, the number of personal pronouns does not and cannot reflect the amount of othering accomplished by explicit references to out-groups, which occurs often across white nationalist fora via a wide range of terms (Baumgarten, 2017). The explicit references outside of personal pronouns were tracked as part of the manual analysis here. Those which referenced out-groups alone accounted for between 2.6 % and 7.3 % of words across the eight authors (note: this number excludes personal pronouns). This means that the othering was even more common than one could have concluded from the LIWC output alone as not all the terms identified are included the LIWC dictionary and those that are may be classified under other headings, including **social** or **culture** words.

This problem of the underrepresentation of othering from using LIWC categories as a proxy is related to the patterns of false negative errors amongst the psychological process categories (and the **verb** category). For the **verb** category (and other linguistic categories) many of the false negatives could be avoided by using a part-of-speech tagger. For the psychological process categories, however, the errors are not as easily remedied. The concern is amplified when one considers that the false negative counts did not include missed implicit meanings (this is outside of LIWC's capabilities and would therefore be an unfair penalty to impose). Thus, the comprehensiveness of the program's account of the data is limited not only by its problems capturing words/phrases which explicitly represent the same construct or share a meaning with other entries in the dictionary, but also by its inability to account for the words/phrases which achieve the same result via implication. Relatedly, it is worth noting that there are instances where false negatives for one of the examined categories are words included within other categories in the dictionary (e.g., *invaded* is not included in the **verb** category but is in **power**). This represents a fundamental flaw in the categorization approach within LIWC where, like with the 'othering' terms, words that represent similar sentiments are spread across a range of categories, making it difficult to see certain functional patterns that would be apparent in discourse analyses. The only remedy to these problems appears to be to carry out a manual discourse analytic approach alongside the LIWC analysis, which reduces the claimed benefit of the program's effectiveness and efficiency that is based on it being able to automatically analyze large textual datasets to extract meaningful functional and psychological patterns.

The implications for the errors that contributed consistently to the observed patterns noted in the results section are many. First, the high false positive rates on some linguistic categories and high false negative rates on some psychological process categories have potential consequences related to the "summary variables". While the algorithms used to calculate these variables are proprietary and themselves unpublished, the LIWC manual points to specific articles which provide basic information on the LIWC categories used in them (Cohn et al., 2004; Kacewicz et al., 2014; Newman et al., 2003; Pennebaker et al., 2014). Assuming these papers do in fact reflect the categories used in the algorithms, the outcomes for all four summary variables would be impacted by the high false positive rates amongst many of the linguistic categories examined in this paper, and the *authenticity* and *emotional tone* outcomes would be further impacted by the high false negative rates in the psychological process categories. *Analytical thinking*, for instance, uses **prepositions** and **auxiliary verbs,** among others (Pennebaker et al., 2014). Given that precision was about 80 % for **prepositions** and **under 50 %** for **auxiliary verbs**, it calls into question the reliability of the algorithms and their outputs. For variables such as *authenticity* (Newman et al., 2003) and *emotional tone* (Cohn et al., 2004), the authors used a combination of linguistic and psychological process categories, including **affect** (and its subordinate **emotion** categories), **cognition,** and **social,** all of which were found to have high false negative rates here (and thus low recall scores). Without knowing the weight assigned to each of these categories in the four algorithms, it is difficult to posit what the extent of the impact of the high error rates would be on the summary variables. However, it is evident from the analysis that caution should be exercised when generating conclusions based on the summary variables when using similar data.

These error patterns undermine the claims of LIWC's creators and proponents that the inability to consider context is unlikely to drastically influence the results, especially with more data. Even though this analysis was conducted with a small dataset, the frequency and consistency of both error types across the dataset suggests that the impact from the disregard for context is not a negligible one and would be unlikely to be resolved with the introduction of more data. Of course, not all the LIWC categories were examined here, but the findings here at the least suggest that more caution is warranted when drawing conclusions and generalizations based on the LIWC output alone, particularly when it comes to the type of data analyzed in forensic and security contexts.

Conclusions

The findings of this study suggest that LIWC cannot be considered reliable, efficient, and effective for the analysis of large online datasets in forensic and security contexts. It may be part of a solution that also involves considerable manual checking and analysis, but it provides an insufficient amount of information on its own, at least with this type of data. More specifically, it may be that LIWC could be used as a signposting tool, but it would be dangerous to use it in either a research or operational context *without devoting proper resources to checking its results and using further qualitative forms of analysis to understand the functional meanings of any identified tokens in context*. This would allow any missed or misclassified tokens to be captured, the relationships between linguistic elements to be observed, and the patterns of language and meaning in a text to be understood more fully, though it would grossly affect the efficiency of LIWC, one of its key selling points.

That an in-depth investigation such as this one had not yet, to our knowledge, been carried out presents a major concern given how widely used LIWC has been in the decades since its conception, and even more so given its expansion into research in forensic and security contexts. In the future, as for any tool, this type of critical examination of LIWC's performance should be conducted on other types of data to determine whether the errors are as common for other LIWC categories and for genres and contexts which are both more similar to those used to create

the dictionary and which differ from online extremist fora. Until such analyses are carried out to surface, evaluate, and mitigate any issues of reliability, LIWC should not be overly or uniquely depended upon for critical applications such as those found in forensic and security contexts.

CRediT authorship contribution statement

Madison Hunter: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Tim Grant:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Madison Hunter reports financial support was provided by UK government. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Anti-Defamation League (2023). *Hate on display™hate symbols database*. https://www.adl.org/resources/hate-symbols/search.

Baele, S.J., 2017. Lone-actor terrorists' emotions and cognition: an evaluation beyond stereotypes. Polit. Psychol. 38 (3), 449–468. https://doi.org/10.1111/pops.12365.

Baker, P., Vessey, R., 2018. A corpus-driven comparison of English and French Islamist extremist texts. Int. J. Corp. Linguist. 23 (3), 255–278. https://doi.org/10.1075/ijcl.17108.bak.

Bakhtin, M.M., 1986. Speech Genres and Other Late Essays. University of Texas Press.

Baumgarten, N., 2017. Othering practice in a right-wing extremist online forum. Language @ Int. 14 (1).

Bednarek, M. (2009). Emotion talk and emotional talk: cognitive and discursive perspectives. In H. Pishwa, *Language and Social cognition: Expression of the Social Mind* (pp. 395–431). De Gruyter.

Bell, A., 1984. Language style as audience design. Lang. Soc. 13 (2), 145–204.

Biber, D., Conrad, S., Reppen, R., 1998. Corpus linguistics: Investigating language Structure and Use. Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., 1999. Longman Grammar of Spoken and Written English. Pearson Education Limited.

Booth, A., 2023a. Collective Identity and Careers in a White Nationalist Forum. Aston University. Doctoral thesis.

Booth, A., 2023b. Fractured in-group identity (re)negotiation in an online white nationalist forum. Appl. Corp. Linguist. 3 (3), 100062. https://doi.org/10.1016/j.acorp.2023.100062.

Boyd, R.L., Ashokkumar, A., Seraj, S., Pennebaker, J.W., 2022. The Development and Psychometric Properties of LIWC-22. University of Texas at Austin. https://www.liwc.app/help/psychometrics-manuals.

Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. Qual. Res. Psychol. 3 (2), 77–101.

Chinchor, N., Sundheim, B.M., 1993. MUC-5 evaluation metrics. In: Proceedings of the 5th conference on message understanding (MUC5 '93). Association for Computational Linguistics. https://doi.org/10.3115/1072017.1072026.

Chung, C.K., Pennebaker, J.W., 2011. Using computerized text analysis to assess threatening communications and behavior. In: Chauvin, C. (Ed.), Threatening Communications and behavior: Perspectives on the Pursuit of Public Figures. National Academies Press.

Chung, C.K., & Pennebaker, J.W. (2018a). Textual analysis. In H. Blanton, J.M. LaCroix, & G.D. Webster, *Measurement in Social Psychology*. Routledge.

Chung, C.K., & Pennebaker, J.W. (2018b). What do we know when we LIWC a person? Text analysis as an assessment tool for traits, personal concerns and life stories. In V. Zeigler-Hill & T.K. Shackelford, *The Sage handbook of Personality and Individual Differences*. Sage.

Cohn, M.A., Mehl, M.R., Pennebaker, J.W., 2004. Linguistic markers of psychological change surrounding September 11, 2001. Psychol. Sci. 15 (10), 687–693. https://doi.org/10.1111/j.0956-7976.2004.00741.x.

Don, A., 2016. It is hard to mesh all this": invoking attitude, persona and argument organisation. Funct. Linguist. 3 (1), 9. https://doi.org/10.1186/s40554-016-0033-1.

Drouin, M., Boyd, R.L., Hancock, J.T., James, A., 2017. Linguistic analysis of chat transcripts from child predator undercover sex stings. J Forens. Psychiatry Psychol. 28 (4), 437–457. https://doi.org/10.1080/14789949.2017.1291707.

DuBois, J.W. (2007). The stance triangle. In Englebretson R., *Stancetaking in discourse: Subjectivity, evaluation, Interaction*. John Benjamins Publishing Company.

Figea, L., Kaati, L., Scrivens, R., 2016. Measuring online affects in a white supremacy forum. In: 2016 IEEE Conference on Intelligence and Security Informatics (ISI). https://doi.org/10.1109/ISI.2016.7745448.

Francis, M.E., Pennebaker, J.W., 1992. Putting stress into words: the impact of writing on physiological, absentee, and self-reported emotional well-being measures. Am. J. Health Promot. 6 (4), 280–287. https://doi.org/10.4278/0890-1171-6.4.280.

Gales, T., 2010. *Ideologies of violence: A corpus and Discourse Analytic Approach to Stance in Threatening Communications* (Publication No. 3422713) [Doctoral dissertation, University of California-Davis]. ProQuest Dissertations Publishing.

Gillings, M., 2024. Corpus Linguistic Approaches to Deception Detection. Taylor & Francis.

Glasgow, K., Schouten, R., 2014. Assessing violence risk in threatening communications. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. https://doi.org/10.3115/v1/W14-3205.

Gottschalk, L.A., Gleser, G.C., Daniels, R.S., Block, S., 1958. The speech patterns of schizophrenic patients: a method of assessing relative degree of personal disorganization and social alienation. J. Nerv. Ment. Dis. 127 (2), 153–166. https://doi.org/10.1097/00005053-195808000-00008.

Grant, T., Baker, K., 2001. Identifying reliable, valid markers of authorship: a response to Chaski. Forensic Linguistics 8, 66–79.

Grover, T., Mark, G., 2019. Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit. In: Thirteenth International AAAI Conference on Web and Social Media. https://doi.org/10.1609/icwsm.v13i01.3221.

Halliday, M.A.K., Matthiessen, C.M.I.M., 2014. *Halliday's Introduction to Functional Grammar* (4th ed.). Routledge.

Hurt, M., 2020. Pledging to harm: A linguistic Analysis of Violent Intent in Threatening Language. Aston University. Doctoral thesis.

Institute for Economics & Peace. (2022). *Global terrorism index 2022: measuring the impact of terrorism*.

Kaati, L., Shrestha, A., Sardella, T., 2016. Identifying warning behaviors of violent lone offenders in written communication. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW). https://doi.org/10.1109/ICDMW.2016.0152.

Kacewicz, E., Pennebaker, J.W., Davis, M., Jeon, M., Graesser, A.C., 2014. Pronoun use reflects standings in social hierarchies. J. Lang. Soc. Psychol. 33 (2), 125–143. https://doi.org/10.1177/0261927×13502654.

Landis, J.R., Koch, G.G., 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 33 (2), 363–374.

Lorenzo-Dus, N., Kinzel, A., 2019. So is your mom as cute as you?': examining patterns of language use by online sexual groomers. J. Corp. Discour. Stud. 2 (1), 1–30.

Martin, J.R., White, P.R.R., 2005. The Language of evaluation: Appraisal in English. Palgrave MacMillan.

Miron, M.S., Douglas, J.E., 1979. Threat analysis: the psycholinguistic approach. FBI Law Enforcem. Bull. 48 (9), 5–10.

Mudde, C., 2019. The Far Right Today. John Wiley & Sons.

Newman, M.L., Pennebaker, J.W., Berry, D.S., Richards, J.M., 2003. Lying words: predicting deception from linguistic styles. Pers. Soc. Psychol. Bull. 29 (5), 665–675. https://doi.org/10.1177/0146167203029005010.

Newsome-Chandler, H., Grant, T., 2024. Developing a resource model of power and authority in anonymous online criminal interactions. Language and Law/Linguagem e Direito 10 (1), 110–130. https://doi.org/10.21747/21833745/lanlaw/10_1/a4.

PennebakerConglomerates. (2025). *How it works*. https://www.liwc.app/help/howitworks.

Pennbaker, J.W., King, L.A., 1999. Linguistic styles: language use as an individual difference. J. Pers. Soc. Psychol. 77 (6), 1296–1312.

Pennebaker, J.W., & Chung, C.K. (2009). Computerized text analysis of Al-Qaeda transcripts. In K. Krippendorff & M.A. Bock, *The Content Analysis Reader*. Sage.

Pennebaker, J.W., Chung, C.K., Frazee, J., Lavergne, G.M., Beaver, D.I., 2014. When small words foretell academic success: the case of college admissions essays. PLoS One 9 (12), e115844. https://doi.org/10.1371/journal.pone.0115844.

Pennebaker, J.W., Booth, R.J., Boyd, R.L., & Francis, M.E. (2015). LIWC2015 Operator's Manual. Pennebaker Conglomerates. https://www.liwc.app/help/psychometrics-manuals.

Read, J., Carroll, J., 2012. Annotating expressions of Appraisal in English. Lang. Resour. Eval. 46 (3), 421–447. https://doi.org/10.1007/s10579-010-9135-7.

Shuy, R.W., 2014. The Language of Murder Cases. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199354832.003.0004.

Swales, J.M., 1990. Genre analysis: English in Academic and Research Settings. Cambridge University Press.

Tausczik, Y.R., Pennebaker, J.W., 2010. The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. 29 (1), 24–54. https://doi.org/10.1177/0261927×09351676.

Tayebi, T., 2016. Exploring the underlying expectations. J. Pragmat. 101, 1–17. https://doi.org/10.1016/j.pragma.2016.05.006.

Torregrosa, J., Thorburn, J., Lara-Cabrera, R., Camacho, D., Trujillo, H.M., 2020. Linguistic analysis of pro-ISIS users on Twitter. Behav. Sci. Terror. Polit. Aggress. 12 (3), 171–185. https://doi.org/10.1080/19434472.2019.1651751.

vanLeeuwen, T., 1996. The representation of social actors. In: Caldas-Coulthard, C., Coulthard, M. (Eds.), Texts and practices: Readings in Critical Discourse Analysis. Routledge, pp. 32–70.

Vickers, P., Barrault, L., Monti, E., Aletras, N., 2024. We need to talk about classification evaluation metrics in NLP. arXiv Preprint. arXiv:2401.03831.