Connecting Peace Research and Natural Language Processing with Cultural Violence for Hostile Narrative Analysis

# Abstract

This article extends a previous article by the authors to rethink aspects of hate speech detection as hostile narrative analysis using Galtung’s theory of cultural violence. Guiding this development is the research question, ‘*to what extent do the current computational methods of NLP understand hostile narratives?*’. The article responds to this question by reviewing quantitative methods in hate speech detection. In addition to problems arising from the polysemy of hate speech itself, the article finds that quantitative methods in NLP have questionable explanatory rigour. The development of hostile narrative analysis, therefore, incorporates more qualitative approaches to NLP. The first qualitative element responds to the polysemy of hate speech by using Galtung’s cultural violence from Peace Research to develop a methodological framework. This framework detects what Galtung refers to as the Self-Other gradient between an orator’s ingroup and their outgroup. This Self-other gradient reflects the extent of ingroup elevation and outgroup othering in violence legitimisation. The underpinning research for hostile narrative analysis draws upon a representative dataset consisting of Mein Kampf by Adolf Hitler, and political texts from President George Bush and Osama bin Laden from the War on Terror. In that he advocated for non-violence, texts from Martin Luther King provide control data. The second qualitative element is in the computational methods of hostile narrative analysis derived from the methodological framework. In contrast to current quantitative approaches, these computational methods parse a text by language clauses using pattern-based NLP. This article develops upon prior work by explaining the text pre-processing requirements for the computational methods. To generate meaningful outputs, text pre-processing must address the abstract representation of outgroups in a hostile narrative. To the best of our knowledge, this is the first attempt to develop cultural violence for NLP, and this combination presents new research opportunities for PeaceTech.

# Introduction

This article responds to the research question, ‘to what extent do the current computational methods of natural language processing (NLP) *understand* hostile narratives?’. Many descriptions of NLP suggest it gives machines the ability to understand natural language. Hirschberg and Manning (2015) suggest, ‘(c)omputational linguistics, also known as natural language processing (NLP), is the subfield of computer science concerned with using computational techniques to learn, *understand* (italics added), and produce human language content’ (Hirschberg and Manning 2015: 261). Bird *et al*. (2009) take NLP ‘in a wide sense to cover any kind of computer manipulation of natural language’, whether simple tasks such as counting word frequencies or ‘*understanding* (italics added) complete human utterances’, to the extent of generating at least ‘useful responses’ (Bird, Klein, and Loper 2009: ix). NLP provides transformative value for functional applications like chatbots and natural language translation. In contrast, this article finds limitations with quantitative methods for more subjectively defined applications like hate speech detection and hostile narrative analysis.

An Association of Computer Machinery (ACM) article for the 2021 Web Science Conference first introduced Hostile narrative analysis (see: XXXX, XXXX). A hostile narrative itself is a story for legitimising violence against another person or group. Underpinning hostile narrative analysis is Galtung’s (1990) theory of cultural violence from Peace Research to understand violence legitimisation. Galtung (1990) explains how cultural violence works ‘by changing the moral colour of a [violent] act from red/wrong to green/right or at least to yellow/acceptable’ (Galtung 1990: 292). To analyse a hostile narrative, therefore, is to reveal the legitimisation of violence in natural language. This article extends XXXX (XXXX) by framing the development of a hostile narrative analysis using quantitative and qualitative methods and makes the following contributions:

1. A critical analysis of hate speech detection (Section 2)
2. A methodological framework of cultural violence to analyse hostile narratives (Section 3)
3. An evaluation of existing NLP technologies for the hostile narrative analysis method (Section 4)

While quantitative methods currently dominate NLP, this article responds to the research question by adopting a mixed methods approach that includes qualitative approaches. A somewhat problematic dichotomy between qualitative and quantitative methods began in the mid-twentieth century. ‘By the 1940s and 50s in sociology, psychology and some other fields, quantitative method (in the form of survey and experimental research) had become the dominant approach’ (Hammersley 1992: 40). ‘The concept *qualitative research* started to spread during the 1950s and 1960s and became widespread in large parts of the social sciences during the 1970s and 1980s but in some disciplines such as psychology it did not gain momentum until the 1980s and 1990s’ (Allwood and Allwood 2012: 1421). The dichotomy is a `mistaken belief that qualitative researchers are in the business of interpreting stories and quantitative researchers are in the business of producing fact’ (boyd and Crawford 2012: 667). Yet, contrary to any claims of objectivity by quantitative researchers, all research tasks – especially the linguistic analysis – have interpretative elements and require a mixed approach of quantitative and qualitative methods.

This article facilitates the inclusion of qualitative approaches in NLP by employing Explainable AI to situate the understanding of a hostile narrative within an explanatory dialogue. In a shift from opaque to more human-understandable algorithms, the Explainable AI movement is gaining increased provenance in the computer sciences (see: Adadi and Berrada 2018; Saeed and Omlin 2021). Central to Explainable AI is an ‘explanatory dialogue’ between an enquirer and explainer through which understanding evolves over time (XXXX (XXXX)). A machine’s role is to generate rigorous inputs to these dialogues. As explained in the next section, the polysemy of hate speech questions the explanatory rigour of current quantitative methods in hate speech detection. In response, the third section begins the qualitative element of hostile narrative analysis with a methodological framework derived from cultural violence theory. This framework rethinks aspects of hate speech detection as hostile narrative analysis. XXXX (XXXX). The second qualitative element, therefore, draws upon pattern-based NLP to generate more rigorous inputs to an explanatory dialogue. The final section then reviews the text pre-processing requirements for this approach. This pre-processing address the abstract representation of outgroups in a text to support any claim of understanding a hostile narrative. The resulting approach incorporates qualitative and quantitative methods to facilitate the human understanding of hostile narratives.

Lindgren (2020) explains how quantitative methods generally analyse data tables with statistical tools, while qualitative methods typically involve close reading of textual data from interviews, observations and documents (Lindgren 2020: 13). The quantitative aspects of NLP are an application of statistical methods. For example, Latent Dirichlet allocation (LDA) is a ‘generative probabilistic model’ of text often used for inferring the topics of a text (Blei, Ng, and Jordan 2003: 996). Text classification (critically analysed in XXXX (XXXX)) is where machine learning algorithms employ statistical models based on [manually annotated] training data (Atteveldt, Velden, and Boukes 2021: 4). ‘Word embedding’ infers meaning from the mathematical relationships of ‘distributed representations of words in a vector space’ (see: Mikolov, Chen, Corrado, et al. 2013; Mikolov, Chen, Sutskever, et al. 2013; Mikolov, Yih, and Zweig 2013). While statistical methods enable large-scale systems development, XXXX (XXXX) and this article show how these methods provide questionable value for more subjectively defined applications like hate speech detection, sentiment analysis and hostile narrative analysis.

This article joins a growing provenance of qualitative approaches in NLP. ‘Qualitative approaches are typically used to explore new phenomena and to capture individuals’ thoughts, feelings, or interpretations of meaning and process’ (Given 2012: xxix). As such, Lindgren (2020) reflects on the growing integration of the social and computer sciences by proposing the idea of ‘Data Theory’ as a `broad label for a hybrid form of digital social science research practice that is data-intensive, computational (‘quantitative’), yet theoretically interpretative (‘qualitative’)’ (Lindgren 2020). To qualitatively assess a text, researchers typically use semiotics, narrative, content, discourse, archival, and phonemic analysis, in addition to statistics, tables, graphs, and numbers (Denzin and Lincoln 2018: 12). The growing field of Digital Humanities sees increasing integration of these tools with computational methods to understand social phenomena (see: Dunn and Schuster 2020; Orlandi 2021; van der Zwaan et al. 2017). The qualitative approach to hostile narrative analysis draws upon Peace Research, Semiotics and Narrative Inquiry to facilitate the human understanding of hostility in natural language.

To the best of our knowledge, the methodological framework presented in this article is the first attempt to theoretically develop cultural violence beyond Galtung’s (1990) paper. Developing hostile narrative analysis using cultural violence seeks to contribute to the emerging field of PeaceTech. The idea of PeaceTech emerged in 2015 with ‘the creation of the United States Institute of Peace’s (USIP) PeaceTech Lab as an umbrella term for a focus on new information and communication technologies and their role in building peace’ (Rhian 2018: 13). As a general observation for PeaceTech, problems with current computational methods arise from a ‘technical first’ approach before applying theory; this paper reverses that approach and begins with cultural violence theory to guide technological development. Accordingly, this article is written from a practitioner’s perspective to inform peace researchers about applying computational methods to peace-building applications; for the technical audience, the contributions seek to explain the value of pattern-based methods for NLP. The following section presents a critical analysis of hate speech detection to establish the requirement for hostile narrative analysis.

# How effective are the current methods of hate speech detection?

The growing hate speech detection industry is attracting significant attention and investment. Many organisations are documenting the increasing awareness of hate speech online (Anon 2020; HOPE not Hate 2022; Tell MAMA 2020). For online platforms, the emerging UK Online Harms Bill will obligate web companies to tackle the propagation of hate on their platforms[[1]](#footnote-2). The UK’s Alan Turing Institute’s website records academic interest across research groups, journals, workshops, and conference sessions[[2]](#footnote-3). For Peace Research, hate speech gains mention in aspects of electoral violence (Fjelde 2020: 142; Smidt 2020), authoritarian governance (Bayer, Bethke, and Lambach 2016: 763) and early warning for genocide (Goldsmith et al. 2013: 450). While attracting such attention, nevertheless, a 2019 Alan Turing Institute review finds ‘the field is beset with terminological, methodological, legal and theoretical challenges’ (Vidgen, Margetts, and Harris 2019:3). In support of this review, this section finds that through an interdisciplinary transition from the legal profession to the social, political and computer sciences, hate speech itself has unhelpfully become a polysemous term. This polysemy then questions the effectiveness of current approaches to hate speech detection.

## Why is hate speech a polysemous term?

The polysemy of hate speech arises from what Brown (2017) observes as its ‘legal’ and ‘ordinary’ meaning. He explains how ‘different kinds of people who are not legislators, legal professionals or scholars of law use the term ‘hate speech’ in countless different types of contexts about a tremendous diversity of phenomena’ (Brown 2017: 424). Matsuda (1989) provides a legal meaning of hate speech to criminalise ’a narrow, explicitly defined class of racist speech, to provide public redress for the most serious harm, while leaving many forms of racist speech to private remedies’ (Matsuda 1989: 2380). The focus of criminalising hate speech was US jurisprudence, and Matsuda was careful to respect the First Amendment rights of Free Speech. In doing so, Matsuda (1989) provides three identifying characteristics to distinguish the most harmful aspects of hate speech from other forms of racist and non-racist speech (Matsuda 1989: 2357):

• The message is of racial inferiority

• The message is directed against a historically oppressed group; and

• The message is persecutorial, hateful and degrading.

The first characteristic is the primary identifier of hate speech, whereby all target group members are at once generalised as homogeneous and inferior. The second recognises structural elements of racism for which racist speech is a mechanism of subordination that reinforces historical injustice. While recognising historical injustice, Matsuda (1989) concedes, ‘should history change course, placing former victim groups in a dominant or equalised position, the newly equalised group will lose the special protection suggested here’ (Matsuda 1989: 2362). The third characteristic recognises the potential of written or spoken words to incite hatred or violence against a target. And where these characteristics focus on racism, Matsuda intended to take a more general view of hate speech for other minority groups.

The ordinary meaning arises from the interdisciplinary transition of hate speech into the social and political sciences, where legal precision gives way to reactionary politics. As Brown (2017) observes,

‘Hate speech’ has been perhaps most often associated with liberal progressives, or people on the left of politics – who use it to highlight and problematise speech that they view as racist, xenophobic, homophobic, Islamophobic, misogynistic, disablist, or in some other way targeted at minority groups in ways that supposedly violate ideals of respect, solidarity, tolerance, and so forth (Brown 2017:425).

Brown (2017) also observes an opposing reactionary position whereby,

political and religious conservatives repudiate such uses of the term and view them simply as crude attempts to close down meaningful debate on what they believe are the evils of open-border policies, the failures of multiculturalism as a social experiment, the lamentable decline of traditional moral values, political correctness gone made, and so on.’ (Brown 2017:425).

For this second position, Mondon and Winter (2020) note that Free Speech has become a ‘reactionary tool’ without any concrete legal basis to push racist agendas (Mondon and Winter 2020: 75–79). With such a range of interpretations of hate speech in the social and political sciences, its meaning depends on who uses the term.

A transition into the computer sciences has only exaggerated the polysemy of hate speech. Following the liberal progressive position identified in Brown (2017), a literature review by Schmidt and Wiegand (2017) characterises hate speech as a ‘broad umbrella term for numerous kinds of insulting user-created content’ and references 18 developer-defined definitions of hate speech, some including the problematic prefix ‘cyber’ (Schmidt and Wiegand 2017: 1). Subsequent literature shows how the absence of a generally accepted definition remains (Abro et al. 2020: 484; Chiril et al. 2022: 323; Fortuna, Soler-Company, and Wanner 2020: 6786; Kosisochukwu, Gao, and Xue 2020: 155; Mullah and W. N. W. M. Zainon 2021: 88366). In response, developers tend to offer an application-specific definition. As a general observation, rather than a definition shaping the computational methods, the computational methods seem to shape the definition.

## How does the polysemy of hate speech affect hate speech detection?

Problems with the polysemy of hate speech are most apparent in the labelling schemas and training datasets used in the quantitative method of text classification from NLP. In an early paper, Maron (1961) explains how text classification (which he refers to as automatic indexing) concerns ‘the problem of deciding automatically what a given document is “about”’ (Maron 1961:404). More technically, ‘if *d*i is a document of the entire set of documents D, and {c1, c2, …, cn} is the set of all the categories, then text classification assigns one category cj to a document di’ (Ikonomakis, Kotsiantis, and Tampakas 2005: 966). In less-technical terms, D might be a blog, social media post or newspaper article, and *d*i is a constituent element of D, whether a paragraph, sentence, phrase or word. As noted in four literature reviews, text classification is the most common computational method of hate speech detection (Fortuna 2018: 22; Kovács, Alonso, and Saini 2021: 4; Mullah and W. M. N. W. Zainon 2021a: 88364; Schmidt and Wiegand 2017: 2).

The quantitative aspect of text classification relies upon pre-labelled training data for classifying natural language inputs. For creating training data, hate speech detection literature generally draws upon far-right and extremist content from websites such as Twitter and Gab (Burnap and Williams 2015a; Nithyanand, Schaffner, and Gill 2017), Reddit (Vidgen et al. 2021) or a combination of website content with offensive word lists (Nithyanand et al. 2017). For classifying the constituent elements, hate speech classifiers most commonly use binary classifications of {hateful, non-hateful} (Burnap and Williams 2015a: 231; Fortuna 2018: 22; Mullah and W. M. N. W. Zainon 2021b: 1; Poletto et al. 2020: 497), or multi-classifications, such as {identity-directed abuse, affiliation-directed abuse, person-directed abuse, non-hateful slurs, counter speech} (Vidgen et al. 2021: 2291). Classifiers then use a range of statistical methods to determine the similarity of an input to a labelled element of the training data. As a simple example, training data containing a hateful annotation for ‘parasite’ would give the same annotation to an input document containing the same term.

Disagreements in developing labelling schemas and training data for annotating training data arise from the polysemy of hate speech and often reflect the contrasting positions observed by Brown (2017). Any determination of hatefulness may depend less on an orator’s intended meaning and more on annotators’ subjective assessment (see: Davidson et al., 2019; Geva et al., 2019; Gonen & Goldberg, 2019; Wiegand et al., 2019). The problem is disagreement in whether document *d*i is indeed hateful in the absence of a generally agreed definition of hate speech. Moreover, binary classifications of {hateful, non-hateful} are likely too reductive for broad interpretations of language. Once the resultant training data are applied to real-world research tasks, ‘definitions of abuse fast become embroiled in contentious debates around privacy, freedom of speech, democracy, discrimination, and the power of big tech companies’ (Vidgen et al. 2020: 10).

The application of quantitative methods to detect such a polysemous concept gives only an illusory understanding of hate speech. In addition to text classification, Fortuna identifies several other commonly used quantitative methods in hate speech detection, such as topic classification, sentiment analysis, text classification and word embeddings (Fortuna 2018: 20). Using such quantitative methods, nonetheless, implies that subjective assessments of hatefulness are somehow quantifiable. If quantitative methods are about objectivity and identifying facts, the quantification of hatefulness is only relevant to a developer's specific definition of hate speech, annotation schema and training data. Consequently, the outputs of detection systems are system-specific, are not generalisable, and provide questionable rigour to explanatory dialogues. As such, the proposal to rethink hate speech as hostile narrative analysis seeks to remedy problems with the polysemy of hate speech by using Galtung’s (1990) more generally accepted theory of cultural violence to develop computational methods of improved explanatory rigour.

# How does Cultural Violence enable the understanding of hostile narratives?

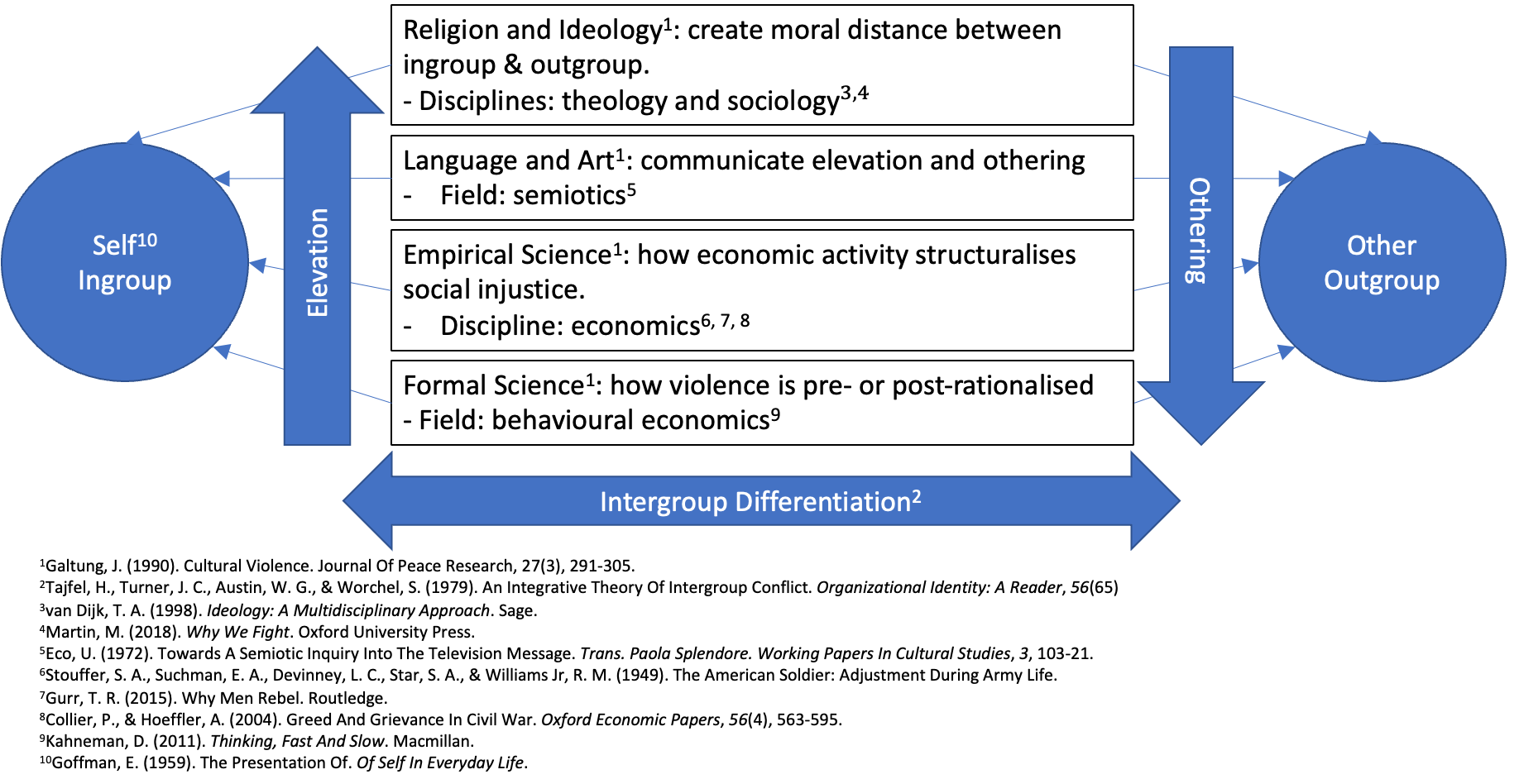


Figure 1. A Theoretical Framework of Cultural Violence

This section develops the cultural violence methodological framework shown in Figure 1 as one way to enable the *human* understanding of a hostile narrative. The methodology is about detecting how an orator may elevate their ingroup while othering their outgroup in violence legitimisation and is underpinned by the following hypothesis (XXXX (XXXX)):

The more significant the intergroup differentiation created by elevation and othering, the more legitimate violence against an outgroup becomes.

The following subsection connects elements of cultural violence to academic literature as a basis for the framework in Figure 1. The subsequent subsection augments the framework with Tajfel’s and Turner’s social identity theory from social psychology (see: Tajfel 1974; Tajfel and Turner 1979). By doing so, cultural violence itself becomes a unifying idea for interdisciplinarity. The subsequent sections then use a method derived from this framework to develop hostile narrative analysis. This framework provides the first qualitative element of hostile narrative analysis, and the emphasis on *human* understanding applies to a machine’s input to an explanatory dialogue.

## What is Cultural Violence?

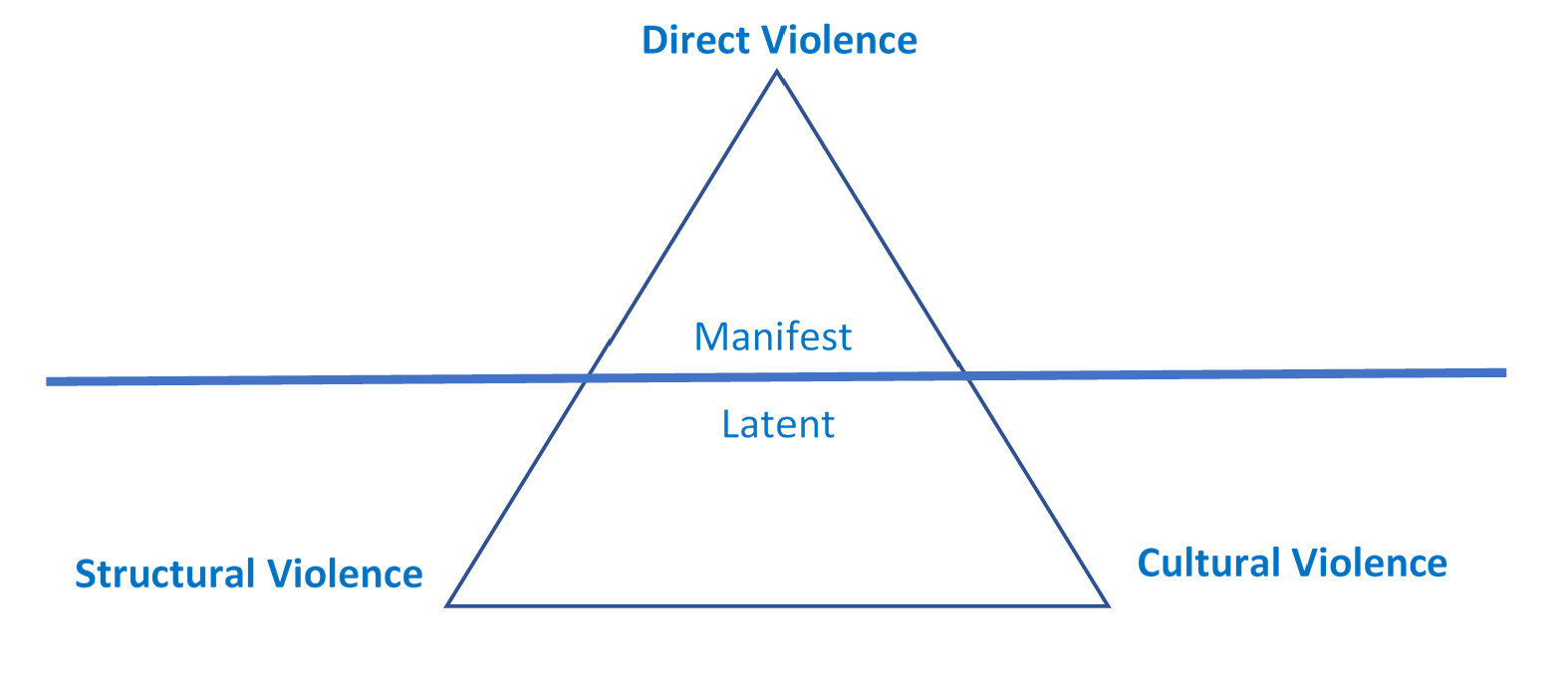


Figure 2. Galtung's Violence Triangle

Galtung (1990) defines cultural violence as:

Those aspects of culture, the symbolic sphere of our existence – exemplified by religion and ideology, language and art, empirical science and formal science (logic, mathematics) – that can be used to justify or legitimise direct or structural violence (Galtung 1990: 291).

Two parts of this definition provide structure for its explanation. For the second part about direct and structural violence, Galtung conceives violence broadly and within a triangular framework shown in Figure 2. Direct, structural, and cultural violence go beyond the common understanding of violence as a physical act (Webel and Galtung 2007: 23). Direct violence is physically harming others with intention; structural violence is the harm arising from socio-political structures and decisions that deprive someone of their access to the basic needs necessary for fulfilling their life potential; cultural violence is the cultural justification of direct and structural violence (Galtung and Fischer 2013: 151). As a complement to Matsuda (1989), therefore, the utterance of speech deemed unlawful is an act of hate speech, which becomes a subclass of direct violence. Cultural violence – the focus of this paper – then represents how an orator might legitimise such violent utterances. In turn, hostile narrative analysis seeks to reveal the legitimisation of structural and direct violence in natural language.

This subsection elaborates on how each aspect of culture – religion and ideology, language and art, and empirical and formal science – features in violence legitimisation by creating a ‘Self-other gradient’ between an ingroup and an outgroup. Galtung’s (1990) Self-other gradient is about how an orator elevates a ‘Chosen People’ (Self) above those deemed ‘lower down the scale of worthiness’ (Other) (Galtung 1990: 302). Elevation and othering align with the types of structural violence associated with hate speech. Matsuda *et al.* (1993) explain racism as the structural subordination of a group based on an idea of racial inferiority and supremacy (Matsuda et al. 1993: 36), while Augoustinos & Reynolds (2001) note how ‘the power one group has over another transforms race prejudice into racism’ (Augoustinos and Reynolds 2001: 4). Accordingly, the methodological framework in Figure 1 uses a range of literature to explain how aspects of each cultural domain create a Self-other gradient through elevation and othering.

The first domains of religion and ideology refer to how divine and secular codes feature in conceptualising the Self-other gradient. Literature on religion and ideology is vast; this explanation draws upon three scholars who summarise Galtung’s ideas. In his theological analysis, Wright (2010) describes scriptures from Christianity, Judaism and Islam as ‘maps of the landscape of religious tolerance and intolerance, maps that amount to a kind of code for the salvation of the world’ (Wright 2010: 322). Equally, for van Dijk (1998), ‘ideologies allow people as group members, to organise the multitude of social beliefs’ about ‘good or bad, right or wrong, for *them*, and to act accordingly’ (Dijk 1998: 8). As Martin (2018) observes, while the moral codes underpinning ideologies emerge from within societies, religion adds belief in an external supernatural deity as the arbitrator of morality (Martin 2018: 163). Whether divine or secular, orators use these codes as cognitive frameworks to create moral difference in favour of the Self and against the other, thereby creating a gradient between each.

The field of Semiotics explains how the third and fourth cultural domains of language signify moral codes in linguistics and visual systems. De Saussure defined language (langue) as ’a system of signs that express ideas’ and introduced Semiotics as a science to ‘investigate the nature of signs and the laws governing them’ (Saussure 1916: 16). Semiotics has since become a theory of meaning (or ‘signification’) that analyses the network of relationships between terms of linguistic systems, referred to as vocabularies, within the minds of a speech community (Knapp et al. 2008: 231). Words like ‘friends’ contrast with such antonyms as ‘enemies’ to categorise the Self and the other. An orator creates a gradient between each by using framing words that signify such ‘moral oppositions’ as ‘good’ and ‘evil’ or ‘right’ and ‘wrong’ (Chandler 2017:114). Accordingly, language and art signify moral codes to communicate the Self-other gradient and are central to analysing hostile narratives.

For the fifth cultural domain, empirical science, Galtung (1990) describes how a ‘law of comparative advantage’ is embedded into economic activity to create status differences. The idea of ‘Relative Deprivation’ (RD), first introduced by Stouffler *et a*l*.* (1949)and elaborated by W. G. Runciman (see: Runciman, 1966; Webber, 2007, 2021), explains how people derive their status in comparison to others. Gurr (2010) later used RD to explain how a decline in status compared to expectations may lead to violence (see: Gurr, 2010). Conversely, in the ‘Greed and Grievance’ debate, Collier and Hoeffler use macroeconomic analysis to show how pursuing status as power and profit can lead to violent rebellion (see: Collier et al., 2009; Collier & Hoeffler, 1998, 2004). Arising from economic activity, therefore, is the potential for violence in response to the perceived decline or pursuit of status.

The sixth cultural domain, formal science, can be explained by prospect theory to understand how people rationalise choices under risk. From 1979, Kahneman and Tversky pioneered this theory to challenge existing economic assumptions giving rise to the field of behavioural economics. Kahneman (2011) subsequently explains how the quick-thinking brain develops heuristics (roughly rules of thumb) to rationalise the world, often at the expense of more considered thinking (Kahneman 2011: 7). Prospect theory has become a rejoinder to the rational choice models of behaviour prevalent in the social sciences (see: Webber, 2021). Arising from heuristics is the potential for biases manifest as discriminatory beliefs about the Self and others. A Self-other gradient itself could be considered a heuristic used by either a person or group to rationalise harm.

Aspects of these six cultural domains of religion and ideology, language and art, and formal and empirical science create a gradient between the Self and Other through elevation and othering. As belief systems, religion and ideology provide cognitive frameworks of moral codes which elevate the Self in contrast to the Other. Semiotics explains how linguistic and visual systems signify moral codes to communicate elevation and othering. Empirical science explains how economic activity may create elevation and othering through perceived status differences. Formal science, as behavioural economics, consolidates the other cultural domains to show the logic of how violence against the Other might be pre- or post-rationalised. Accordingly, the steeper the gradient between the Self and Other created by elevation and othering, the more legitimate direct or structural harm becomes. With the basis for cultural violence established, the following subsection elaborates on the Self and other using social identity theory.

## How does Social Identity Theory Augment Cultural Violence Theory?

Social identity theory augments cultural violence theory to understand groups and group formation in violence legitimisation. As the originators of social identity theory, Tajfel and Turner (1979) describe groups as:

…a collection of individuals who perceive themselves to be members of the same social category, share some emotional involvement in a common definition of themselves and achieve some degree of social consensus about the evaluation of their group and their membership of it’ (Tajfel and Turner 1979: 283).

For group membership, Goffman (1959) theorises how individuals gain their sense of Self through their ‘performance’ in different group settings (Goffman 1959: 8). The process of violence legitimisation then begins with the representation of the Self and others as members of differentiated ingroups and outgroups.

The process of ‘social categorisation’ leads to group formation and the abstract representation of ingroups and outgroups in natural language. As Hogg (2008) observes, ‘categorisation renders the world more predictable’, thus facilitating planning for effective action (Hogg 2008: 74). When social categories become salient, people see themselves and others less as individuals and more as homogenous group members. Social groups then take on the persona of a prototypical group member. The prototype is not an objective reality but rather a ‘subjective sense of the defining attributes of a social category that fluctuates according to context’ (Hornsey 2008: 208). The group prototypes are then given such abstract representations in natural language as ‘friend’ or ‘enemy’ to signify the ingroup and outgroup. In how abstract representations feature in violence legitimisation, ‘when the other is not only dehumanised but has been successfully converted into an ‘it’, deprived of [humanity], the stage is set for any type of direct violence’ (Galtung 1990: 298).

As Hogg (2016) explains, ‘intergroup differentiation’ suggests people are concerned with ensuring their ingroup is positively distinctive, clearly differentiated from and more favourably evaluated than identified outgroups (Hogg 2016). Differentiation often manifests when people define themselves by who they are *not* rather than who they are; ‘a sense of identity is founded upon a distinction between us and the rest of the world’ (Chandler 2017: 108–14). Turner and Reynolds (2011) subsequently found that having self-identified with a group, individuals tend to assign more resources to their ingroup than their outgroup (Turner and Reynolds 2011). As Tajfel and Turner (1979) note, in contrast to ethnocentric notions of group formation, experiments yield ‘the basic and highly reliable finding’ that ‘trivial, ad hoc categorisation’, such as shared scores on a maths test, ‘leads to ingroup favouritism and discrimination against the outgroup’ (Tajfel and Turner 1979: 39).

While differentiation is a reoccurring theme in theories of intergroup relations, hate speech detection research seems only concerned with othering (see: Alorainy et al., 2018; Cao et al., 2020; Fortuna, 2018; Poletto et al., 2020; Schmidt & Wiegand, 2017). To only consider othering further questions the explanatory rigour of hate speech detection systems. In contrast, the cultural violence methodological framework is connected to established literature to provide a theoretical basis for explanatory dialogues about hostile narratives. As such, this framework represents how a *human* understands violence legitimisation in natural language. The subsequent sections draw upon a method derived from this framework to assess the computational aspects of understanding.

# To what extent do current NLP methods enable the understanding of hostile narratives?

The following section is about the computational methods of hostile narrative analysis that are derived from the methodological framework in Figure 1. This section begins by introducing the method derived from this framework and the corpus to represent hostile narratives. This method and corpus provide the means to assess NLP algorithms for hostile narrative analysis. The section then extends upon the pattern-based approach introduced in XXXX (XXXX) by reviewing the text pre-processing requirements for this method that particularly account for the abstract representation of outgroups in a text.

## What is the method for analysing hostile narratives?

|  |  |  |
| --- | --- | --- |
| Objective | Description | Output |
| Obj 1. | Text Pre-processing | Labelled data |
| Obj 2. | Detect and classify named entities as either ingroup or outgroup | A list of named entities classified as either ingroup or outgroup |
| Obj 3. | Detect and classify ingroup elevation and outgroup othering phrases. | A list of elevation and othering statements linked to each entity |
| Obj 4. | Score the Self-other gradient. | Scores for each ingroup and outgroup relationship of a text |

Figure 3. The method for analysing hostile narratives

XXXX (XXXX) first introduced the hostile narrative analysis method shown in Figure 3. The experiments for developing the computational methods of hostile narrative analysis are organised around each objective and are available online[[3]](#footnote-4).

* Objective 1 is primarily about text pre-processing for the subsequent objectives. These pre-processing tasks generally draw upon the quantitative method of text classification to label words by their grammatical, semantic, and syntactic properties.
* Objective 2 is about detecting social categorisation in natural language whereby an orator categorises named entities according to whether they are an ingroup or outgroup. For NLP, ‘named entities are definite noun phrases that refer to specific types of individuals, such as organisations, persons, dates, and so on’ (Bird et al. 2009: 281).
* Objective 3 then seeks to detect the clauses an orator uses to create a Self-other gradient between the groups identified in objective 2. The central feature of these clauses is the use of religion and ideology to create moral difference.
* Objective 4 relies upon a quantitative method to score the Self-other gradient. The words associated with a group in a clause are scored to generate a value for elevation and othering.

XXXX (XXXX) reviews the more qualitative approach of pattern-based NLP for objectives 2 and 3, while this section focuses on text pre-processing. The outputs of each objective then provide inputs to a rigorous explanatory dialogue and promote a human understanding of hostile narratives

## Introducing the hostile narrative corpus

As stories for legitimising violence against people or groups, the research underpinning this article represents hostile narratives with historical texts that have featured in violence legitimisation. In narratology terms, each text represents a different *genre* of hostile narrative. Hitler’s *Mein Kampf* represents the antisemitic genre; declarations of war by President George Bush and Osama bin Laden from the ‘War on Terror’ represent the warfare genre. In that he advocated for non-violence, speeches by Martin Luther King provide control data. Each hostile narrative genre then comprises different dialects comprising distinct vocabularies of terms, and warfare terminology likely features in most other genres as a metaphor. Each orator’s intention in these texts is generally accepted, therefore, they provide a stable reference and avoid the contentious debates to which Vidgen *et al.* (2020) and Brown (2017) refer. This article focuses on the warfare genre with Bush’s declaration of war in response to the 9/11 attacks in New York in 2001.

Bush’s dataset provides examples of the abstract representation of named entities with noun phrases or pronouns to represent ingroups and outgroups. As Galtung (1990) explains, using noun phrases or pronouns to represent groups is a process of othering. Hitler described the Jews as the ‘dangerous it’, the ‘vermin’, or ‘bacteria’; Stalin described the ‘kulaks’ in political terms as the ‘class enemy; Reagan described Qadhafi as the ‘mad dog’; Washington experts describe ‘terrorists’ as the ‘cranky criminals’ (Galtung 1990: 298). In their response to the outgroup, ingroup members conversely elevate themselves as the ‘heroes’ or ‘saviours’ of a story. This article shows how the abstract representation of named entities is a particular feature of a hostile narrative that NLP must address.

**Noun Phrases**

Our friends

Our great country

Our nation

Our people

Our communities

Fellow

Brethren

The Ummah

Brothers/sisters

Sons/daughters

Martyrs

**First Person & First Person Plural Pronouns**

I

Me

My

Mine

We

Us

Our

Ourselves

**Nouns Phrases**

Enemy

Murderer

Oppressor

Crusader

Terrorist

Terrorist organisation

Terrorist parasites

You

Your

Yours

Yourself

s/he

Him/her

His/her

Hers

Herself

Himself

It

Its

Itself

They

Them

Their

themselves

**3rd thing, 2nd & 3rd Person Pronouns**

**InGroup Phrases**

**Outgroup Phrases**

Figure 4. Pronouns and Noun Phrases of Bush’s and bin Laden’s texts.

Figure 4 shows a sample of noun phrases and pronouns used by Bush and bin Laden to refer to their ingroup and outgroups. As the figure shows, first-person pronouns such as ‘we’ or ‘our’ refer to an ingroup, whereas second and third-person pronouns like ‘you’ or ‘they’ can refer to either an ingroup or outgroup. Familial noun phrases like ‘brothers’ and ‘sisters’ or ‘sons’ and ‘daughters’ signify an ingroup, while phrases like ‘crusader’ or ‘terrorist’ signify an outgroup. These noun phrases have varying degrees of intensity for elevation and othering and for scoring the Self-other gradient. As will be shown throughout the rest of this paper, any understanding of a hostile narrative should link these abstract representations to the named entities to which they refer.

## What are the text pre-processing requirements to address the abstract representation of groups in a hostile narrative?

|  |  |  |  |
| --- | --- | --- | --- |
| Objective | Description | Technology | Output |
| Pre-processing | | | |
| Obj 1.1 | Tokenisation - splitting a text into meaningful segments known as a token | spaCy tokeniser | A spaCy Doc object containing tokens |
| Obj 1.2 | Pos Tagging - Assign part-of-speech tags. | spaCy tagger | Tokens labelled with pos tags |
| Obj 1.3 | Parser – assign dependency labels | spaCy dependency parser er | Tokens with dependency labels |
| Obj 1.4 | Lemmatizer  – add a word lemma to each token | spaCy lemmatiser | Tokens with added lemmas as labels |
| Obj 1.5 | Named entity recognition and disambiguation  – detect and label named entities | spaCy named entity recognition  spaCy entity linker | Tokens with assigned entity labels |
| Obj 1.6 | Named Concept Recognition – label the concepts of a text | Custom component | Tokens with assigned concept labels |
| Obj 1.7 | Co-reference resolution  – assign named entities to pronouns and noun phrases. |  | Modified text by resolving noun phrases to named entities. |

Figure 7. The Computational Method of Each Objective

Thus far, this paper has sought to establish the requirement for more qualitative approaches in NLP to analyse a hostile narrative. The second and fourth sections explain how quantifying meaning from a text provides questionable inputs to an explanatory dialogue about hostile narratives. The third section establishes more qualitative aspects using a methodological framework of cultural violence and the hostile narrative analysis method. XXXX (XXXX) shows how the more qualitative approach of pattern-based NLP improves upon quantitative methods by generating a rationale to explain why an output is produced. This fifth section now reviews the text pre-processing requirements in objective 1 to enable the pattern-based NLP of objectives 2 and 3.

As this section explains, the pre-processing requirements must address the abstract representation of groups in a text to reveal the target of violence in a hostile narrative. Figure 6 shows the pre-processing requirements identified thus far. The computational methods for each objective draw upon the spaCy NLP python library, which provides many production-ready components and the ability to create bespoke add-ins. The following explanation uses five sample sentences from Bush’s declaration of war to explain how pre-processing addresses abstract representation.

Sentence 1: The evidence we have gathered all points to a collection of loosely affiliated terrorist organisations known as al Qaeda.

Sentence 2: This group and its leader -- a person named Usama bin Laden -- are linked to many other organisations in different countries, including the Egyptian Islamic Jihad and the Islamic Movement of Uzbekistan.

Sentence 3: The leadership of al Qaeda has great influence in Afghanistan and supports the Taliban regime in controlling most of that country

Sentence 4: They hate our freedoms

Sentence 5: These terrorists are the heirs of all the murderous ideologies of the 20th century

These sentences are spread across Bush’s text and refer to the outgroups against whom he declares war. The first three sentences identify Bush’s outgroups, which apply to objective 2. The first sentence establishes ‘al Qaeda’ as a terrorist organisation, and the second refers to them as ‘this group’. The second and third sentences establish additional outgroups by linking ‘the Egyptian Islamic Jihad’, ‘the Islamic Movement of Uzbekistan’ and ‘the Taliban Regime’ to ‘al Qaeda’. The fourth and fifth sentences are othering statements and apply to objective 3. ‘They’ or ‘these terrorists’ in each sentence collectively signify ‘al Qaeda’, ‘the Taliban Regime’, ‘the Egyptian Islamic Jihad’ and ‘the Islamic Movement of Uzbekistan’.

The following clauses of interest in the first three sentences identify Bush’s outgroups for objective 1:

Sentence 1: The evidence we have gathered all points to {a collection of loosely affiliated terrorist organisations}subject {known as}predicate al {Qaeda}object

Sentence 2, Clause 1: {This group}subject {is linked to}predicate {many other organisations in different countries}object.

Sentence 2, Clause 2: {Its leader}subject {is linked to}predicate {many other organisations in different countries}object.

Sentence 2, Clause 3 {its leader}subject {--}predicate  {a person named Usama bin Laden}object

Sentence 2, Clause 4: {many other organisations in different countries}subject, {including}predicate {the Egyptian Islamic Jihad}object and {the Islamic Movement of Uzbekistan}object

Sentence 3, Clause 1: {The leadership of al Qaeda}subject  {has}predicate1 {great influence in Afghanistan}object1 and {supports}predicate2 {the Taliban regime}object2 in controlling most of that country

A human would subconsciously resolve the named entity of ‘al Qaeda’ to the noun phrase ‘this group’ and pronoun ‘its’, and then resolve the named entity ‘Usama bin Laden’ to the noun phrase, ‘[al Qaeda’s] leader’. Similarly, ‘the leadership of al Qaeda’ also includes ‘Usama bin Laden’. ‘These terrorists’ and ‘they’ in the fourth and fifth sentences refer to this collection of outgroups. The relevant clauses from each sentence, therefore, subconsciously resolve as follows:

Sentence 1, Clause 1: a collection of loosely affiliated terrorist organisations named al Qaeda

Sentence 1, Clause 4: many other organisations in different countries, include the Egyptian Islamic Jihad and the Islamic Movement of Uzbekistan[[4]](#footnote-5)

Sentence 2, Clause 1: al Qaeda is linked to the Egyptian Islamic Jihad and the Islamic Movement of Uzbekistan.

Sentence 2, Clause 2: Usama bin Laden is linked to the Egyptian Islamic Jihad and the Islamic Movement of Uzbekistan.

Sentence 2, Clause 3: Usama bin Laden is al Qaeda’s leader

Sentence 3, Clause 1: The leadership al Qaeda supports the Taliban regime

Sentence 4, Clause 1: al Qaeda, the Egyptian Islamic Jihad, the Islamic Movement of Uzbekistan, and the Taliban regime hate our freedoms

Sentence 5, Clause 1: al Qaeda, the Egyptian Islamic Jihad, the Islamic Movement of Uzbekistan, and the Taliban regime are the heirs of all the murderous ideologies of the 20th century

A human interpretation of these sentences resolves the abstract terms to the named entities to which they refer; any NLP claiming to understand a hostile narrative must do the same. Without the necessary pre-processing steps, therefore, any text subsequently processed by an NLP algorithm is a misrepresentation of an orator’s intended meaning. The following subsections summarise experiments with spaCy components for the pre-processing steps shown in Figure 6. Each experiment assesses how well they contribute to reproducing each sentence to the human interpretation given above.

## Objective 1.1 – 1.4: Semantic Labelling

Objectives 1.1 to 1.4 are about labelling the semantic units of a text. spaCy provides pre-trained language models in a text classification architecture for text pre-processing. As a functional task, spaCy’s algorithm uses statistical models and word embeddings to predict the attributes of each word. There are 23 models for different world languages[[5]](#footnote-6); the English model used here uses the OntoNotes 5, a large, annotated corpus of 2.9 million words comprising telephone conversations, newswire, newsgroups, broadcast news, broadcast conversation, weblogs, religious texts (Weischedel et al. 2012). spaCy augments these texts with WordNet, which is a lexical database of English words. Of note, the majority of OntoNotes texts are from news and broadcast news, therefore, this model strongly represents the geopolitical vocabulary of the hostile narrative corpus. Other genres and dialects require more bespoke models.



Figure 8. The Semantic Units of Sentences 1 and 2

Figure 8 depicts the semantic units of sentences 1 and 3 whereby spaCy labels each word as either a noun, proper noun, pronoun or verb[[6]](#footnote-7). For Objective 1.1, tokenisation is the process of splitting text into minimal meaningful units such as words, punctuation marks [and] symbols’ (Patel and Arasanipalai 2021: 13). For Objective 1.2, part-of-speech tagging is about assigning an attribute label to each token to indicate a word’s grammatical function in a sentence (Bengfort, Bilbro, and Ojeda 2018:44). For objective 1.3, dependency labelling assigns attributes to each token according to a word’s syntactic function in a sentence. Lemmatization refers to representing a word using its conical head, known as a lemma, and is used in NLP to generalise words by their different tenses or spellings. The noun chunks shown in Figure 8 are multi-word noun phrases that abstractly represent named entities. The spaCy components for these tasks operate to state-of-the-art standards and are not assessed further here.

## Objective 1.6: Named Entity Recognition and Resolution

Named entity recognition (NER) and resolution enable objective 1 to detect groups in a text for subsequent classification as an ingroup or outgroup. Named entity recognition is the `task of classifying tokens of interest in a sequence of tokens into specific entity types, such as a person, an organisation, or a location’ (Patel and Arasanipalai 2021: 62). Entity resolution is about assigning a unique identifier to different mentions of the same named entity across a text. Named entity resolution means an individual ID identifies the same entity. For example, ‘the United States of America’ as a named entity is also represented by the terms ‘USA’, ‘US’ and the ‘United States’, therefore, each should attract the same unique ID.

The second experiment assesses the accuracy of spaCy’s language model for NER and a spaCy add-on for entity resolution[[7]](#footnote-8). spaCy labels named entities using the OntoNotes 5 NER schema, which has 18 entity types (Weischedel et al. 2012). spaCy’s classification algorithm uses contextual information about a word to assign the label. As such, the algorithm differentiates the named entity ‘US’ from the pronoun ‘us’. Entity-Fishing[[8]](#footnote-9) is a spaCy add-on for the entity fishing tool[[9]](#footnote-10) for named entity resolution. The algorithm uses Wikidata[[10]](#footnote-11) to provide an ID that uniquely identifies a named entity. Of particular utility for explanatory dialogues, entity-fishing also provides a confidence score for the ID, the associated Wikidata description for the named entity, and alternate IDs. These additional outputs offer a rationale within an explanatory dialogue for assessment by human judgement.

spaCy and Entity-Fishing detect and resolve the named entities from the test sentences as follows:

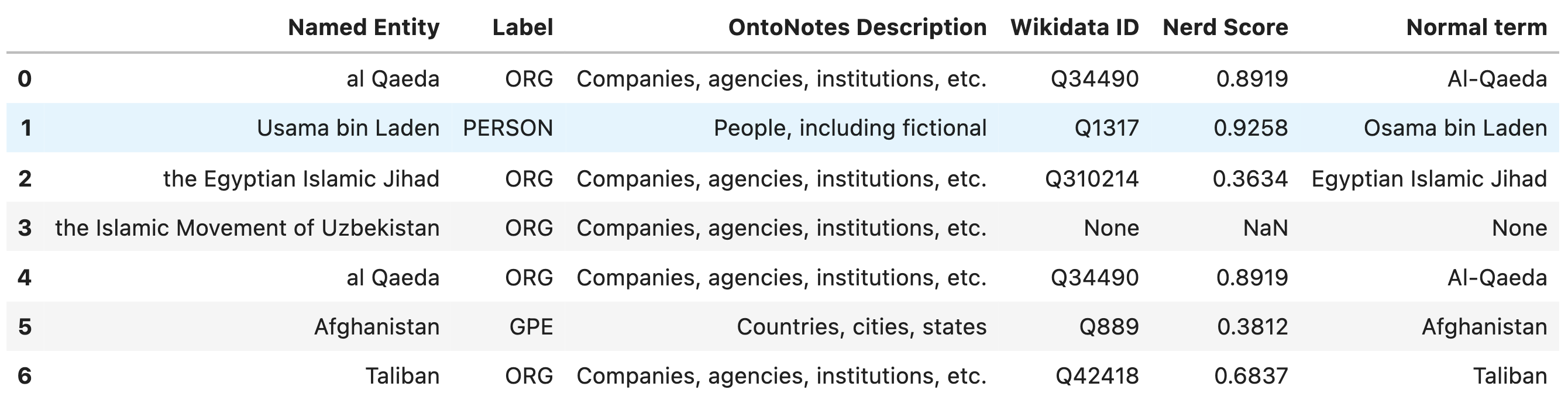


Figure . The named entities and Wikidata IDs for the test sentences.

The named entity column shows the original text. The label and description column each show the spaCy label and Ontonotes description. Note how the labels do not introduce biases, for example, ‘al Qaeda’ is labelled as an organisation and not a terrorist group. The Wikidata ID column shows the unique identifier from Wikidata, while the Nerd Score is the confidence level for that identifier. The ‘normal term’ is the normative reference for the original text. While Wikidata recognises ‘Usama bin Laden’ and ‘the Egyptian Islamic Jihad’, it does not recognise ‘the Islamic Movement of Uzbekistan’. A review evaluated the label assigned to each named entity using external sources, and an Islamic academic was consulted for bin Laden’s text to ensure accuracy. In response, a custom NER component was created from these corrections, with 65 corrections for bin Laden and 31 for Bush[[11]](#footnote-12).

### Objective 1.5 - Named Concept Recognition

Under the idea of named concept recognition, the research underpinning this article is developing a schema for labelling a word’s meaning. Named concept recognition draws upon quantitative coding to label concepts associated with religion and ideology in a text. As a quantitative process, ‘codes are commonly created prior to data collection…concepts and hypotheses are most often developed in advance, and categories and their codes are derived deductively from theory or borrowed from the extant literature’ (Benaquisto 2008: 85). The schema groups synonymous concepts from Bush and bin Laden’s speeches into eight different group contexts: social, medical, academic, religious, political, economic, security and military[[12]](#footnote-13). XXXX (XXXX) shows the utility of named concept recognition for classifying language clauses and for generating rationales for an explanatory dialogue. As such, named concept recognition enables the identification of ingroups and outgroups in objective 2 and the identification of elevation and othering in objective 3.

## Objective 1.7: Co-Reference Resolution

While the preceding objectives prepare data for processing, coreference resolution directly addresses abstract representations. Coreference resolution is about identifying all the pronouns and noun phrases that refer to the same named entity, as explained in the introduction to this section. For objective 2, coreference resolution assists with identifying the noun phrase an orator uses to represent named entities. For objective 3, coreference resolution detects the elevation and othering statements that refer to a named entity in the subject of a clause

The following assessment summarises the performance of three coreference resolution algorithms for the spaCy pipeline. The first algorithm is ‘coreferee’ from explosion.ai; the second is ‘neuralcoref’, developed by Huggingface[[13]](#footnote-14); the third is ‘coref’, provided by AllenNLP[[14]](#footnote-15). While ‘neuralcoref’ draws upon neural networks only, both ‘coref’ and ‘coreferee’ also include a mixture of rules and machine learning. The assessment reviews each algorithm’s output using the example sentences above to assess how well they resolve the noun phrases referring to the groups in Bush’s text.

Graphical user interface, text, application, email

Description automatically generated

Figure 10. Coreference outputs of sentences 1 and 2.

Figure 8 summarises the assessment with each algorithm's coreference outputs for sentences 1, 2, 4 and 5. Most notably, Coreferee incorrectly resolves ‘they’ in sentence 4 with ‘Americans’. In context, the section this sentence is drawn from barely mentions Bush’s outgroups, however, ‘Americans’ is most proximate to ‘they’. Six sentences separate sentences 1 and 2 with only one other mention of ‘al Qaeda’ between them. Coreferee fails to make this connection and adequately resolve sentence 2. Coref, on the other hand, makes the connection between ‘al Qaeda’ and the pronoun ‘its’ and produces a plausible output. AllenNLP successfully connects both ‘this group’ and ‘its’ to the entire noun phrase from sentence 1, ’a collection of loosely affiliated terrorist organisations known as al Qaeda’. While technically a correct representation of how a human would interpret the sentence, further processing is required to recreate the resolved sentence using ‘al Qaeda’ and not ‘a collection of loosely affiliated terrorist organisations known as al Qaeda’.

While the AllenNLP and Coref algorithms generate plausible outputs, this assessment shows how these algorithms are limited by the number of entities they can resolve to a noun phrase. Where a human subconsciously resolves the four mentioned organisations to ‘they’ and ‘these terrorists’, coreference algorithms only link a single reference. The algorithms work well for shorter texts, but manual intervention through an explanatory dialogue is required to make these links over a longer narrative.

# Discussion

This research responds to the question of whether NLP can understand hostile narratives by seeking to incorporate qualitative elements into NLP. Responding to the question finds limitations with quantitative methods for hate speech detection. The polysemy of hate speech creates a disconnect between an explanatory theory and computational methods, thus giving rise to two problems. Firstly, the absence of a defining theory leads to unrepresentative and contentious training datasets. Secondly, the quantitative nature of these methods inappropriately attempts to quantify a qualitative problem.

Consequently, the outputs of a detection system are specific to a developer’s bespoke definition of hate speech. The outputs of a hate speech detection system, therefore, provide questionable rigour to an explanatory dialogue. The response is to rethink hate speech detection as hostile narrative analysis using cultural violence.

This article presents the methodological framework of cultural violence as a way to understand hostile narratives. To summarise this framework: social identity theory explains group formation and sources of intergroup differentiation, while cultural violence explains how a Self-other gradient rationalises differentiation and legitimises violence between groups. A hostile narrative itself communicates the religious and ideological belief systems an orator uses to create moral distance through elevation and othering. The method derived from this framework informs an explanatory dialogue about how elevation and othering contribute to the legitimisation of violence. This framework has general applicability to different hostile narrative genres about which hate speech is generally concerned. For example, orators may frame their ingroup and outgroups by attributes of race for the racist genre, gender or sex for the sexist or transphobic genres, and sexuality for the homophobic genre.

The mixed method approach to the computational methods of hostile narrative analysis draws upon pattern-based NLP to parse the language clauses of a text. The quantitative methods draw upon machine learning for text pre-processing to label a word's grammatical, semantic and syntactic properties. The qualitative aspect firstly draws upon cultural violence theory as a basis for an explanation. Pattern-based NLP parses language clauses to detect an ingroup and outgroup and how they are elevated and othered across a narrative. And where this use of language clauses enables the production of a rationale, which analysts may then accept, reject, or modify, thereby permitting a rigorous explanatory dialogue about hostile narratives. This paper has additionally demonstrated the level of text pre-processing requirements to address the abstract representation of groups in a text, which is a particular feature of hostile narratives. Nonetheless, human intervention through an explanatory dialogue - the guiding principle of the Explainable AI movement - questions whether NLP can or even should understand a hostile narrative. In answering the research question, therefore, NLP *assists* with the *human* understanding of hostile narratives.

# Conclusion

The research underpinning this paper continues to develop hostile narrative analysis. The primary task is to create test data from the hostile narrative dataset to ensure the computational methods are grounded in empirical observations of violence legitimisation. For continued development, the following areas are of particular interest: developing syntactic patterns, the semiotics of named concept recognition, the scoring schema, visualising outputs, and criticism of hostile narrative analysis. Future research will seek to apply the narrative structures of warfare to other hostile genres like racism. The method, data sources, concept schema, and experiments are available in open source to enable reproducible research for such subsequent research. With this approach and by combining Peace Research with NLP, this research seeks to provide a solid basis for reciprocal collaboration between peace research practitioners and technical developers to create meaningful technologies for PeaceTech.

# Bibliography

Abro, Sindhu, Sarang Shaikh, Zafar Ali, Sajid Khan, Ghulam Mujtaba, and Zahid Hussain Khand. 2020. ‘Automatic Hate Speech Detection Using Machine Learning: A Comparative Study’. *IJACSA) International Journal of Advanced Computer Science and Applications* 11(8):484–91.

Adadi, Amina, and Mohammed Berrada. 2018. ‘Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)’. *IEEE Access* 6:52138–60. doi: 10.1109/ACCESS.2018.2870052.

Allwood, Carl Martin, and C M Allwood. 2012. ‘The Distinction between Qualitative and Quantitative Research Methods Is Problematic’. *Qual Quant* 46:1417–29. doi: 10.1007/s11135-011-9455-8.

Alorainy, Wafa, Pete Burnap, Han Liu, and Matthew L. Williams. 2018. ‘The Enemy Among Us’: Detecting Cyber Hate Speech with Threats-Based Othering Language Embeddings’. *ACM Transactions on the Web* 13(3):39.

Anon. 2020. *AI Research Study:Online Abuse and Project Restart*.

Atteveldt, Wouter van, Mariken A. C. G. van der Velden, and Mark Boukes. 2021. ‘The Validity of Sentiment Analysis:Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms’. *Communication Methods and Measures* 1–20. doi: 10.1080/19312458.2020.1869198.

Augoustinos, Martha., and Katherine J. Reynolds. 2001. ‘Understanding Prejudice, Racism, and Social Conflict’. 362.

Bayer, Markus, Felix S. Bethke, and Daniel Lambach. 2016. ‘The Democratic Dividend of Nonviolent Resistance’. *Journal of Peace Research* 53(6):758–71. doi: 10.1177/0022343316658090.

Benaquisto, Lucia. 2008. ‘Thematic Coding and Analysis’. Pp. 85–88 in *The Sage Encyclopedia of Qualitative Research Methods*. Vol. Volumes 1 & 2, edited by L. M. Given. Sage.

Bengfort, Benjamin, Rebecca. Bilbro, and Tony. Ojeda. 2018. *Applied Text Analysis with Python*. O’Reilly Media, Inc.

Bird, Steven., Ewan. Klein, and Edward. Loper. 2009. *Natural Language Processing with Python*. O’Reilly.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. ‘Latent Dirichlet Allocation’. *Journal of Machine Learning Research* 3(4–5):993–1022. doi: 10.1016/b978-0-12-411519-4.00006-9.

boyd, danah, and Kate Crawford. 2012. ‘Critical Questions for Big Data’. *Information, Communication & Society* 15(5):662–79. doi: 10.1080/1369118X.2012.678878.

Brown, Alexander. 2017. ‘What Is Hate Speech? Part 1: The Myth of Hate’. *Law and Philosophy* 36(4):419–68. doi: 10.1007/s10982-017-9297-1.

Burnap, Pete, and Matthew L. Williams. 2015a. ‘Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making’. *Policy and Internet* 7(2):223–42. doi: 10.1002/poi3.85.

Burnap, Pete, and Matthew L. Williams. 2015b. ‘Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making’. *Policy and Internet* 7(2):223–42. doi: 10.1002/poi3.85.

Cao, Rui, Roy Ka Wei Lee, and Tuan Anh Hoang. 2020. ‘DeepHate: Hate Speech Detection via Multi-Faceted Text Representations’. Pp. 11–20 in *WebSci 2020 - Proceedings of the 12th ACM Conference on Web Science*. Association for Computing Machinery, Inc.

Chandler, Daniel. 2017. *Semiotics: The Basics*. 3rd ed.

Chiril, Patricia, Endang Wahyu Pamungkas, Farah Benamara, Véronique Moriceau, and Viviana Patti. 2022. ‘Emotionally Informed Hate Speech Detection: A Multi-Target Perspective’. *Cognitive Computation* 14(1):322–52. doi: 10.1007/s12559-021-09862-5.

Collier, Paul, and Anke Hoeffler. 1998. ‘On Economic Causes of Civil War’. *Oxford Economic Papers* 50:563–73.

Collier, Paul, and Anke Hoeffler. 2004. ‘Greed and Grievance in Civil War’. *Oxford Economic Papers* 56:563–95. doi: 10.1093/oep/gpf064.

Collier, Paul, Anke Hoeffler, and Dominic Rohner. 2009. ‘Beyond Greed and Grievance: Feasibility and Civil War’. *Oxford Economic Papers* 61(1):1–27. doi: 10.1093/oep/gpn029.

Davidson, Thomas, Debasmita Bhattacharya, and Ingmar Weber. 2019. ‘Racial Bias in Hate Speech and Abusive Language Detection Datasets’. Pp. 25–35 in *Proceedings of the Third Workshop on Abusive Language Online*. Florence: Association for Computational Linguistics.

Denzin, Norman K., and Yvonna S. Lincoln. 2018. *Introduction: The Discipline and Practice of Qualitative Research*.

Dijk, Teun A. van. 1998. *Ideology: A Multidisciplinary Approach*. Sage.

Dunn, Stuart, and Kristen Schuster. 2020. *Research Methods in the Digital Humanities: General Introduction*. edited by K. Schuster and S. Dunn. Routledge.

Fjelde, Hanne. 2020. ‘Political Party Strength and Electoral Violence’. *Journal of Peace Research* 57(1):140–55. doi: 10.1177/0022343319885177.

Fortuna, Paula. 2018. ‘A Survey on Automatic Detection of Hate Speech in Text’. *ACM Computing Surveys* 51. doi: 10.1145/3232676.

Fortuna, Paula, Juan Soler-Company, and Leo Wanner. 2020. ‘Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets’. Pp. 6786–94 in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. Marseille, .

Galtung, Johan. 1990. *Cultural Violence*. Vol. 27.

Galtung, Johan, and Dietrich Fischer. 2013. ‘Violence: Direct, Structural and Cultural’. Pp. 35–40 in *Johan Galtung: Pioneer of Peace Research*.

Geva, Mor, Yoav Goldberg, and Jonathan Berant. 2019. *Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets*.

Given, Lisa. 2012. *The SAGE Encyclopedia of Qualitative Research Methods*. Vol. Volumes 1 & 2. edited by L. M. Given. Sage.

Goffman, Erving. 1959. *Presentation of Self in Everyday Life*. New York: Doubleday Anchor Books.

Goldsmith, Benjamin E., Charles R. Butcher, Dimitri Semenovich, and Arcot Sowmya. 2013. ‘Forecasting the Onset of Genocide and Politicide: Annual out-of-Sample Forecasts on a Global Dataset, 1988-2003’. *Journal of Peace Research* 50(4):437–52. doi: 10.1177/0022343313484167.

Gonen, Hila, and Yoav Goldberg. 2019. ‘Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings but Do Not Remove Them’. Pp. 609–14 in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. Vol. 1.

Gurr, Ted Robert. 2010. *Why Men Rebel*. Paradigm Publishers.

Hammersley, Martyn. 1992. ‘Deconstructing the Qualitative-Quantitative Divide’. Pp. 39–55 in *Mixing Methods: qualitative and quantitative research*. Routledge.

Hirschberg, Julia, and Christopher D. Manning. 2015. ‘Advances in Natural Language Processing’. *Science Magazine* 349(6245):261–66.

Hogg, Michael A. 2008. ‘Social Categorization, Depersonalization, and Group Behavior’. Pp. 56–85 in *Blackwell Handbook of Social Psychology: Group Processes*. Oxford, UK: Blackwell Publishers Ltd.

Hogg, Michael A. 2016. ‘Social Identity Theory’. Pp. 3–17 in.

HOPE not Hate. 2022. *State of Hate 2022*. London.

Hornsey, Matthew J. 2008. ‘Social Identity Theory and Self-Categorization Theory: A Historical Review’. *Social and Personality Psychology Compass* 2(1):204–22. doi: 10.1111/j.1751-9004.2007.00066.x.

Ikonomakis, M., S. Kotsiantis, and V. Tampakas. 2005. ‘Text Classification Using Machine Learning Techniques’. *WSEAS Transactions on Computers* 4(8):966–74.

Kahneman, Daniel. 2011. *Thinking Fast and Slow*. London: Penguin.

Knapp, Karlfried, Gerd Antos Volume, Mouton de Gruyter, · Berlin, and New York. 2008. *Handbooks of Applied Linguistics Communication Competence Language and Communication Problems Practical Solutions*.

Kosisochukwu, Judith Madukwe, Xiaoying Gao, and Bing Xue. 2020. ‘In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets’. Pp. 150–61 in *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics.

Kovács, György, Pedro Alonso, and Rajkumar Saini. 2021. ‘Challenges of Hate Speech Detection in Social Media’. *SN Computer Science* 2(2). doi: 10.1007/s42979-021-00457-3.

Lindgren, Simon. 2020. *Data Theory : Interpretive Sociology and Computational Methods*. Polity.

Maron, M. E. 1961. ‘Automatic Indexing: An Experimental Inquiry’. *Journal of the ACM* 8(3):404–17.

Martin, Mike. 2018. *Why We Fight*. London: Hurst Publishers.

Matsuda, Mari J. 1989. ‘Public Response to Racist Speech: Considering the Victim’s Story’. *Michigan Law Review* 87(8). doi: 10.2307/1289306.

Matsuda, Mari J., Charles R. Lawrence, Richard Delgado, and Kimberlè Williams Crenshaw. 1993. *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. ‘Efficient Estimation of Word Representations in Vector Space’.

Mikolov, Tomas, Kai Chen, Ilya Sutskever, Greg Corrado, and Jeffrey Dean. 2013. ‘Distributed Representations of Words and Phrases and Their Compositionality’. doi: 10.1162/jmlr.2003.3.4-5.951.

Mikolov, Tomas, Wen-Tau Yih, and Geoffrey Zweig. 2013. ‘Linguistic Regularities in Continuous Space Word Representations’. Pp. 746–51 in *Proceedings of NAACL-HLT 2013*. Association for Computational Linguistics.

Mondon, Aurelien, and Aaron Winter. 2020. *Reactionary Democracy: How Racism and the Populist Far Right Became Mainstream*. London: Verso.

Mullah, Nanlir Sallau, and Wan Mohd Nazmee Wan Zainon. 2021a. ‘Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review’. *IEEE Access* 9:88364–76. doi: 10.1109/ACCESS.2021.3089515.

Mullah, Nanlir Sallau, and Wan Mohd Nazmee Wan Zainon. 2021b. ‘Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review | Enhanced Reader’. *IEEE Access* 88364–76. Retrieved 28 January 2022 (https://ieeexplore.ieee.org/document/9455353).

Mullah, Nanlir Sallau, and Wan Nazmee Wan Mohd Zainon. 2021. ‘Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review’. *IEEE Access Nanlir S* 9:88364–76. doi: 10.1109/ACCESS.2021.3089515.

Nithyanand, Rishab, Brian Schaffner, and Phillipa Gill. 2017. ‘Measuring Offensive Speech in Online Political Discourse’.

Orlandi, Tito. 2021. ‘Reflections on the Development of Digital Humanities’. *Digital Scholarship in the Humanities* 36(Supplement\_2):ii222–29. doi: 10.1093/llc/fqaa048.

Patel, Ankur, and Ajay Arasanipalai. 2021. *Applied Natural Language Processing in the Enterprise: Teaching Machines to Read, Write, and Understand*. O’Reilly Media, Inc, USA.

Poletto, Fabio, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. ‘Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review’. *Language Resources and Evaluation 2020 55:2* 55(2):477–523. doi: 10.1007/S10579-020-09502-8.

Rhian, Jennifer Gaskell. 2018. ‘Peacetech Practices and Their Potentials for Empowerment, Participation and Peace’. PhD, University of Southampton.

Runciman, W. G. (Walter Garrison). 1966. ‘Relative Deprivation and Social Justice : A Study of Attitudes to Social Inequality in Twentieth-Century England’. 338.

Saeed, Waddah, and Christian Omlin. 2021. ‘Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities’.

Saussure, Ferdinand de. 1916. *Course in General Linguistics*.

Schmidt, Anna, and Michael Wiegand. 2017. *A Survey on Hate Speech Detection Using Natural Language Processing*. Association for Computational Linguistics.

Smidt, Hannah. 2020. ‘Mitigating Election Violence Locally: UN Peacekeepers’ Election-Education Campaigns in Côte d’Ivoire’. *Journal of Peace Research* 57(1):199–216. doi: 10.1177/0022343319884993.

Tajfel, H., and J. C. Turner. 1979. ‘An Integrative Theory of Intergroup Conflict’. Pp. 33–47 in *The Social Psychology of Intergroup Relations*, edited by W. G. Austin and S. Worschel. Monterey: Brooks/Cole.

Tajfel, Henri. 1974. ‘Social Identity and Intergroup Behaviour’. *Social Science Information* 13(2):65–93. doi: 10.1177/053901847401300204.

Tell MAMA. 2020. *The Impact of the Christchurch Terror Attack*.

Turner, John C., and Katherine J. Reynolds. 2011. ‘Self-Categorization Theory’. Pp. 399–417 in *Handbook of Theories of Social Psychology*. SAGE Publications Ltd.

Vidgen, Bertie, Helen Margetts, and Alex Harris. 2019. *How Much Online Abuse Is There? A Systematic Review of Evidence for the UK Policy Briefing – Full Report*. London.

Vidgen, Bertie, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. ‘Introducing CAD: The Contextual Abuse Dataset’. Pp. 2289–2303 in *2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Vidgen, Bertie, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2020. *Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection*.

Webber, Craig. 2007. ‘Revaluating Relative Deprivation Theory’. *Theoretical Criminology* 11(1). doi: 10.1177/1362480607072737.

Webber, Craig. 2021. ‘Rediscovering the Relative Deprivation and Crime Debate: Tracking Its Fortunes from Left Realism to the Precariat’. *Critical Criminology*. doi: 10.1007/s10612-021-09554-4.

Webel, Charles, and Johan Galtung. 2007. *Handbook of Peace and Conflict Studies*.

Weischedel, Ralph, Sameer Pradhan, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Nianwen Xue, Martha Palmer, Jena D. Hwang, Claire Bonial, Jinho Choi, Aous Mansouri, Maha Foster, Abdel-Aati Hawwary, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, and Ann Houston. 2012. *OntoNotes Release 5.0*.

Wiegand, Michael, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. ‘Detection of Abusive Language: The Problem of Biased Datasets’. Pp. 602–8 in *NAACL-HLT*. Minneapolis: Association for Computational Linguistics.

Wright, Robert. 2010. *The Evolution of God: The Origins Of Our Beliefs*. Little Brown Book Group.

van der Zwaan, Janneke M., Wouter Smink, Anneke Sools, M. Sools Nl, Bernard Veldkamp, Sytske Wiegersma, and Wiegersma Nl. 2017. ‘Flexible NLP Pipelines for Digital Humanities Research’. Utrecht: 4th Digital Humanities Benelux Conference 2017.

1. [Online Harms Whitepaper](https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper), retrieved on 12th Oct 2022 [↑](#footnote-ref-2)
2. [Online Hate Research Hub](https://www.turing.ac.uk/research/research-programmes/public-policy/online-hate-research-hub#:~:text=The%20Alan%20Turing%20Institute.%20The%20%27Hate%20speech%3A%20Measures,social%20media%20and%20news%20platforms.%20University%20of%20Leicester), retrieved on 12th Oct 2022 [↑](#footnote-ref-3)
3. Anning *et al* (2022) [Hostile Narrative Analysis GitHub Repository](https://github.com/Fourthought/Hostile-Narrative-Analysis), retrieved on 12th Oct 2022 [↑](#footnote-ref-4)
4. This clause is re-ordered to assist with the explanation. [↑](#footnote-ref-5)
5. Explosion.ai (2022) [Language Support](https://spacy.io/usage/models#languages), retrieved on 12th Oct 2022 [↑](#footnote-ref-6)
6. Anning (2022) [Semantic Labelling](https://github.com/Fourthought/Hostile-Narrative-Analysis/tree/master/Obj%201%20-%20pre-processing/Experiment%201.1%20-%20Semantic%20labelling), retrieved on 12th Oct 2022 [↑](#footnote-ref-7)
7. Anning (2021) [Named Entity Recognition](https://github.com/Fourthought/Hostile-Narrative-Analysis/tree/master/Obj%201%20-%20pre-processing/Experiment%201.5%20-%20Named%20Entity%20Recognition), retrieved on 12th Oct 2022 [↑](#footnote-ref-8)
8. Luccaterre (2022) [spacyfishing](https://github.com/Lucaterre/spacyfishing), retrieved on 12th Oct 2022 [↑](#footnote-ref-9)
9. Yas1994 (2022) [entity-fishing](https://github.com/kermitt2/entity-fishing), retrieved on 12th Oct 2022 [↑](#footnote-ref-10)
10. Wikimedia Foundation (n.d.) [wikidata](https://www.wikidata.org/wiki/Wikidata:Introduction), retrieved on 12th Oct 2022 [↑](#footnote-ref-11)
11. Anning (2022) [Named Entity Corrections](https://github.com/Fourthought/Hostile-Narrative-Analysis/blob/master/Obj%201%20-%20pre-processing/Experiment%201.5%20-%20Named%20Entity%20Recognition%20/Named%20Entity%20Corrections.ipynb), retrieved on 12th Oct 2022 [↑](#footnote-ref-12)
12. Anning (2022) [Named Concept Recognition](https://github.com/Fourthought/Hostile-Narrative-Analysis/tree/master/Obj%201%20-%20pre-processing/Experiment%201.6%20-%20Named%20Concept%20Recognition), retrieved on 12th Oct 2022 [↑](#footnote-ref-13)
13. HuggingFace (2021) [neuralcoref](file:///Users/stephenanningcorp/Library/Containers/com.microsoft.Word/Data/Library/Preferences/AutoRecovery/NeuralCoref%20is%20a%20pipeline%20extension%20for%20spaCy%202.1+%20which%20annotates%20and%20resolves%20coreference%20clusters%20using%20a%20neural%20network.), retrieved on 12th Oct 2022 [↑](#footnote-ref-14)
14. AllenNLP (2021) [coref](https://github.com/allenai/allennlp-models), retrieved on 12th Oct 2022 [↑](#footnote-ref-15)