British Journal of Political Science (2024), **54**, 179–200
doi:10.1017/S0007123423000042

British Journal of Political Science

ARTICLE

'Super-Unsupervised' Classification for Labelling Text: Online Political Hostility as an Illustration

Stig Hebbelstrup Rye Rasmussen*, Alexander Bor, Mathias Osmundsen and Michael Bang Petersen

Political Science, Aarhus University, Copenhagen, Denmark
*Corresponding author. E-mail: stighj@hotmail.com

(Received 13 February 2022; revised 7 January 2023; accepted 27 January 2023; first published online 24 April 2023)

Abstract

We live in a world of text. Yet the sheer magnitude of social media data, coupled with a need to measure complex psychological constructs, has made this important source of data difficult to use. Researchers often engage in costly hand coding of thousands of texts using supervised techniques or rely on unsupervised techniques where the measurement of predefined constructs is difficult. We propose a novel approach that we call 'super-unsupervised' learning and demonstrate its usefulness by measuring the psychologically complex construct of online political hostility based on a large corpus of tweets. This approach accomplishes the feat by combining the best features of supervised and unsupervised learning techniques: measurements of complex psychological constructs without a single labelled data source. We first outline the approach before conducting a diverse series of tests that include: (i) face validity, (ii) convergent and discriminant validity, (iii) criterion validity, (iv) external validity, and (v) ecological validity.

Keywords: natural language processing; supervised learning; unsupervised learning; social media

As our lives move toward the world of online social media networks, concerns are mounting over the detrimental effects of online hostility on constructive democratic conversations. Studies have demonstrated that people perceive online debates to be hostile and dominated by extreme viewpoints and, as consequence, many are motivated to withdraw (Bor and Petersen 2022). Yet, to better understand these dynamics, researchers are faced with a critical challenge: how to define and measure the important phenomenon of online hostility. The problem is not trivial. As Siegel (2020, 59) concludes in her comprehensive review of the literature on online hostility, the 'absence of clear and consistent definitions' implies that 'our knowledge of the causes, consequences, and effective means of combating' problematic forms of online content 'remains somewhat clouded by definitional ambiguity' and limits 'consensus with regard to the most effective way to detect it across diverse platforms'. In practice, these issues have forced most researchers to turn to a crude 'binary classification task' of documents; for example, hostile or not, toxic or not, and so forth (Siegel 2020). An essential part of the definitional challenge is that hostility is in the eye of the beholder and, in this case, the users. Accordingly, perceptions of hostility may differ from group to group (Rasmussen 2022). For example, some social media users may find that calls for 'cancelling' specific accounts are a form of hostility while other users may view such calls as a way to protect minority voices. Similarly, perceptions of hostile content may change over time. For example, it was only after the 'Black Lives Matter' movement emerged that posts stating 'All Lives Matter' were seen by some as hostile.

https://doi.org/10.1017/S0007123423000042 Published online by Cambridge University Press

Researchers' problems with defining and measuring online political hostility are, in part, a reflection of the context-sensitive nature of hostility.

In this manuscript, we propose a new approach to the study of online political hostility that is designed to capture hostility as it is perceived and used by social media users themselves. This involves two interlinked shifts in the study of online political hostility, one conceptual and one methodological. The first is a conceptual shift towards perceived hostility, or what we will refer to as hostility-in-use rather than exclusive reliance on *a priori* and precisely defined constructs as in, for example, legal definitions of 'hate speech'. Reliance on clearly defined constructs is critically important for some research purposes but they may not exhaust the universe of online content that users find hostile, hence the content that demotivates participation in online spaces. Second, to understand hostility-in-use, a methodological approach is required that can cost-effectively measure such hostility in an era of near-infinite amounts of online text on social media. The key purpose of this manuscript is to outline and empirically validate such an approach.

The methodological approach that we advance to measure hostility-in-use differs from prior research on online political hostility. This research has followed one of three main paths for measuring online hostility as it is expressed through text (for example, comments and posts, tweets and retweets).

The first one is based on a supervised text classification paradigm (Cranmer and Desmarais 2017). It starts with a meticulous hand-labelling of a large batch of texts into 'hostile' and 'not hostile'. These hand-labelled texts then serve as the foundation of a classifier trained to 'learn' and predict whether new texts are hostile or not. This approach thus seeks to generalize from a partially labelled corpus to the rest of a corpus. A prime example of this supervised approach can be found in Djuric et al. (2015), who used word vectors as features to classify online comments into the categories 'clean' and 'abusive'. Another relevant example of such a supervised example is the novel study by Theocharis et al. (2020), studying incivility on Twitter.

The second approach is based on an unsupervised text classification paradigm, such as topic modelling, where 'topics' are not known beforehand but are learned along the way (Blei, Ng, and Jordan 2003; Grimmer 2010; Roberts et al. 2013). Unlike supervised classification, this approach does not require pre-labelled data. Instead, it starts with the texts themselves, attempts to label them externally, and then uses word co-occurrences to uncover which hateful themes and topics are present in the set of texts. For instance, Törnberg and Törnberg (2016) use this approach to study online hostile discussions focusing on Muslims in Sweden.

The third and final approach is based on dictionaries, like sentiment dictionaries (Stine 2019) or specially curated dictionaries of swear words. These dictionaries are then used to score or filter the texts based on a predefined dictionary. A good example of this approach is the study by Rodriguez, Argueta, and Chen (2019) who use sentiment and emotion dictionaries to study hate speech on Facebook. Incivility is also often discussed in the context of dictionaries (Muddiman, McGregor, and Stroud 2019), and we discuss incivility and online hostility in more detail in Appendix 3.1.

Unquestionably, all three approaches to text classification have merits and have cross-fertilized and inspired each other (see, for example, Siegel et al. 2021). However, as we explain below, each approach faces unique challenges that weaken the ability to measure online hostility; in particular, as perceived by the users themselves. We argue that the ideal measurement of political hostility should be an approach (1) that can measure complex constructs at low cost, (2) that reflects users' understanding of political hostility, and (3) that can include contextual information.

In this manuscript, we offer a novel solution to these trade-offs. Our approach makes use of so-called word embeddings to identify what social media users themselves find politically hostile. Critically, however, it goes further than typical word embedding analyses in that it uses these to construct a labelled data set for the entire corpus of text without a single

https://doi.org/10.1017/S0007123423000042 Published online by Cambridge University Press

hand-labelled text.[1] For example, typical applications of embeddings use those embeddings directly to track changes in the relative positions of words over time (Garg et al. 2018) or as input to a classifier (Djuric et al. 2015). We call this method a 'super-unsupervised text classification' technique to signify the use of elements from current supervised and unsupervised approaches to study online political hostility; henceforth, we will simply refer to it as the SU approach.

This manuscript consists of two major steps. First, we outline a set of general principles for the ideal measurement of political-hostility-in-use and discuss how best to live up to them. Second, to validate our approach, we rely on a unique research design combining survey data with scraped behavioural data from a diverse sample of Twitter users. This distinctive research design allows us to investigate: (i) face validity, (ii) convergent and discriminant validity, (iii) criterion validity, (iv) external validity, and (v) the ecological validity of our approach to demonstrate its advantages over current state-of-the-art methods. Importantly, these data sources are only important for validating this approach. The approach itself only requires text data, hence it can be utilized for the entire range of text data available to political scientists studying online hostility. As we discuss, however, this approach is not ideal for all research purposes. It is specifically designed to capture hostility-in-use and not *a priori* defined constructs such as illegal 'hate speech'.

The Ideal Measurement of Online Political Hostility

The SU approach takes seriously what the text producers, or Twitter users in this case, have to say about the world and its constructs. When studying how citizens discuss pro-choice vs pro-life, how politicians discuss democracy, or how Twitter users discuss 'political hostility', rigid preconceived theoretical notions may miss many important distinctions made by users. This is also clear concerning online hostility on Twitter in the US where different ideological camps can hold widely different views on what is right and wrong (Rasmussen 2022). Not taking what users have to say about online hostility, by simply defining it in advance using a specific theoretical definition, is not only very difficult but also misses key aspects of what users perceive to be hostile. In this sense, we can talk of hostility-in-use since what is 'hostile' is defined by the users themselves. On this basis, an ideal approach to measuring perceived online political hostility and other complex psychological constructs is thus characterized by three key features:

(1) It can measure complex constructs, such as online political hostility, at a low cost.
(2) It can measure constructs that reflect the online users' understanding rather than the raters' understanding.
(3) It can include contextual information, such as how constructs vary across subgroups or over time.

Admittedly, while these features are impossible to satisfy fully they are virtues to aspire to. Nonetheless, we believe they are important to highlight since these standards are often not discussed explicitly and, more importantly, current approaches to studying political hostility fall short compared to the method we are suggesting here. We will discuss each feature in turn and explain how current approaches attempt to deal with them. After this, we outline our preferred approach and explain how it aims to comply with these standards.

Measuring complex Constructs at a low Cost?

Researchers taking a hypothetico-deductive approach mostly care about constructs grounded in theory or conceptual work. From this perspective, the ability to classify and categorize a corpus of

[1] The approach thus resembles Watanabe (2021)'s Latent Semantic Analysis (LSA) but, as we discuss in more detail below, there are key conceptual and methodological differences between our suggested approach and the LSA approach.

https://doi.org/10.1017/S0007123423000042 Published online by Cambridge University Press

data is essential to the study of online hostility. In supervised text classification, this feat is often accomplished by teaching raters how to manually 'label' texts according to predefined categories such as 'hate speech', 'offensive language', and 'neither' (see, for example, Davidson et al. 2017). A closely related strategy for text categorization uses dictionaries to score or filter texts (Rodriguez, Argueta, and Chen 2019; Siegel et al. 2021). Unsupervised approaches, like the one taken by Törnberg and Törnberg (2016), cannot directly classify texts based on *a priori* relevant constructs, such as whether a text is politically hostile or not. Using topic modelling to classify tweets into 'politically hostile' and 'not politically hostile', researchers would have to rely on the topic that comes closest to this external labelling; most likely, no single topic or, perhaps, no topic at all would directly fit this externally defined categorization.

One of the main problems with current approaches to measuring externally defined constructs is cost. Hiring student assistants or persuading other researchers to hand-label text is financially costly and exceedingly time-consuming. It takes time and effort to design the coding scheme, repeatedly refine it to meet issues that will likely arise, and read and code large chunks of text. Furthermore, these costs limit the complexity of the constructs that are typically measured. In principle, a coding scheme can be used to classify documents along multiple dimensions, but the tedious work involved in hand coding text means that the typical application has to fall back on a single dimension such as 'hostile' or 'not hostile'. This problem is pertinent when using data from social media where texts often come in millions, and sometimes billions. The hand-labelling of even a small percentage of these is a massive undertaking.

Using Measures That Reflect Online Users' Understanding Rather Than the Raters' Understanding?

Measures of psychological constructs should be ecologically valid. In our case, this implies that texts coded as 'hateful' are perceived as such by social media users themselves. Constructing ecologically valid measures is difficult without grounding the classification process in users' own experiences. This problem is especially pertinent to studying a phenomenon such as online hostility. Beliefs about what hostile speech is is likely to fluctuate across individuals and national contexts (Siegel 2020). Failing to start from the users' understanding of the construct of political hostility also risks relying on the (subjective) and potentially biased labels of external raters. For instance, Wich, Bauer, and Groh (2020) recently demonstrated that the political bias of raters systematically impacted the classification of hate speech. Relatedly, the high costs associated with hand-labelling (see Section: 'Measuring complex constructs at a low cost?') means that the number of raters per text is typically low, ranging from one to four (Barberá et al. 2021; Davidson et al. 2017). While internal reliability can be investigated by reporting inter-rater reliability coefficients, this approach still rests on the work of only a few people's informed but subjective evaluations of whether, say, a given tweet is politically hostile. As researchers and their raters may reflect a narrow subset of the population in terms of racial, sociodemographic, or political diversity (for example, Duarte et al. 2015), this may bias ratings, even when ratings are internally consistent. When studies present classifications of hate and hate speech based on the work of a few raters without disclaiming the ideological background of these same raters, we are left uncertain about the amount of bias present in the labelling exercise.

The topic modelling approach used by Törnberg and Törnberg (2016) is based on the logic that if two words co-occur often in a topic, it is because the respondents themselves perceive that these words tend to go together. Thus, the topics do not necessarily reflect the researcher's perception of the content of a construct but, rather, the respondents' subjective perceptions. Using external coders, such as in the study by Davidson et al. (2017), where the authors label a dataset or by using an externally defined dictionary provides no such reassurance, which is a good reason to prefer the approach taken by Törnberg and Törnberg (2016).

https://doi.org/10.1017/S0007123423000042 Published online by Cambridge University Press

Can Contextual Information be Included When Studying Online Hostility?

The final ideal to aspire to when measuring online political hostility concerns context. Researchers studying online hostility are often interested in comparing their constructs across political groupings – for example, do Democrats and Republicans share common beliefs about what is hostile? – or, perhaps, over time. Is the construct of political hostility even the same across political groupings? Preliminary studies by Wich, Bauer, and Groh (2020) suggest it is not since raters are biased in labelling across the political spectrum. Furthermore, researchers of online hostility could be interested in comparing macro-level factors of online hostility such as inequality or political polarization, which is very difficult to do using dictionaries or hand-labelling.

Incorporating contextual information in the topic modelling approach used by Törnberg and Törnberg (2016) is fully possible. In structural topic modelling, for example, the topic prevalence or topic content (Roberts et al. 2013) can be modelled to examine whether certain topics are more prevalent among particular groups, such as Democrats when compared to Republicans. Conversely, it is difficult to incorporate contextual information in the supervised classification approach of, for example, Davidson et al. (2017) and Djuric et al. (2015). If implicit biases (Wich, Bauer, and Groh 2020) make it difficult to classify online hostility within a country, studying online hostility across countries would seem a Herculean task. And how do we account for measurement invariance; that is, making sure that the constructs we are comparing across contexts are even the same? If we were relying on a latent construct based on survey data, we could of course test for measurement invariance (Meredith 1993). To our knowledge, however, such tests or procedures have not been developed for the hand-labelling of constructs across contexts for text data.

Super-Unsupervised Text Classification

Current approaches to studying online hostility each have distinct advantages and disadvantages. Our contribution is to demonstrate that the use of so-called 'word embeddings' and the choice of focal words offer a simple and cost-effective, yet elegant, solution to the problems of defining and measuring complex constructs like political-hostility-in-use. Since the method draws on both supervised and unsupervised techniques, we label it a 'super-unsupervised' approach. Consistent with an increasing number of political science studies (Rheault and Cochrane 2020; Rodman 2020), we use word embeddings to identify what social media users themselves find politically hostile – that is, hostility-in-use – but, critically, we go further than typical word embedding analyses by using these to label the entire corpus of text.

Specifically, the 'super-unsupervised' technique involves six simple steps: (i) preprocess the text data; (ii) train word vectors for the desired corpus of text (for example, a collection of tweets and retweets); (iii) choose a construct (for example, 'political hostility'); (iv) average the word vectors contained in each document in the corpus (for example, a tweet or retweet) and calculate the distance to the focal construct; (v) sort by the calculated distance; and (vi) appreciate the final 'labelled' dataset.[2] This six-step procedure is illustrated in Fig. 1. In a typical classification setting where the word vectors are used as features for a classifier (for example Djuric et al. (2015)), only the first two steps are used. For those who wish to use the SU approach to study new constructs, we outline a series of general steps a researcher can take and provide some general guidelines to follow in Appendix 5.

Researchers have started using word embeddings in important studies; for example, predicting the ideological placements of political parties (Rheault and Cochrane 2020) and investigating the

[2] More specifically we use the word 'political' and 'hate'. We prefer the word 'hate' over 'hostility' since this word is used more often in everyday language, which our method relies on, compared to the more technical term 'hostility'. A second reason for choosing to call our measure for 'online political hostility' rather than 'online political hate' is to avoid conflating it with the more specific term 'hate speech', which we are not purporting to measure. As discussed below, and illustrated in Fig. 2 in Appendix 3 the method is quite robust to the specific choice of words.

https://doi.org/10.1017/S0007123423000042 Published online by Cambridge University Press

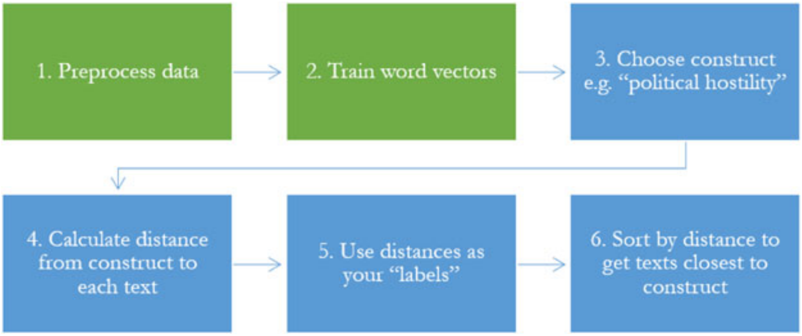
1. Preprocess data
2. Train word vectors
3. Choose construct e.g. "political hostility"
4. Calculate distance from construct to each text
5. Use distances as your "labels"
6. Sort by distance to get texts closest to construct

Figure 1. Illustration of the super-unsupervised approach. The colours refer to how the proposed approach differs from traditional approaches using word embeddings. The 'green' boxes are very typical when using word embeddings, whereas the blue boxes reflect the novel contribution of the super-unsupervised approach.

changing meaning of political concepts over time (Rodman 2020). Accordingly, we are certainly not the first to use word embeddings to study political texts, nor are we claiming to be the first to use word embeddings to classify hostile tweets (for example, Djuric et al. 2015).

However, the present contribution is that by following the approach outlined in Fig. 1, word embeddings can be an exceptionally powerful tool for labelling even extremely large datasets. To our knowledge, no one has used word embeddings to accomplish this goal before. Moreover, we suggest that using this technique will bring us closer to satisfying the three standards outlined previously. Table 1 summarizes the advantages of the super-unsupervised approach compared to traditional approaches to measuring online hostility. We will briefly elaborate on each element in turn.

Super-Unsupervised Learning and the Three Standards for Measurement and Definition of Online Hostility

The super-unsupervised approach utilizes the so-called word2vec methodology and an in-depth understanding of the approach requires an understanding of both this methodology and of language itself. Briefly, the word2vec methodology rests on the so-called 'distributional hypothesis', as explained by Firth (1957): 'You shall know a word by the company it keeps.' The organizing idea is that the meaning of words is not entirely defined by some logical structure of the language itself (*langue*) but rather in its actual usage and everyday understanding (*parole*) (De Saussure 2011).

The idea of using natural language and written words to conduct serious measurements is not as counter-intuitive as it first might appear. This method propelled the widespread success of the five-factor model of personality (that is, the Big Five personality traits). The key idea is that if a personality trait is salient – that is, it is an important and enduring feature in a society – it will be captured by specific words (Matthews, Deary, and Whiteman 2003, Chapter 1). Thus, by

Table 1. Overview of similarities and differences between dictionary-based, topic modelling based and the super-unsupervised approach towards measuring and defining political hostility

	Dictionary	Topic modelling	Classification	Super-unsupervised
Able to measure complex constructs?	Yes	No	Yes	Yes
Constructs reflect the user's self-understanding and not the understanding of the raters	No	Yes	No	Yes
Can include contextual information?	No	No	No	Yes

https://doi.org/10.1017/S0007123423000042 Published online by Cambridge University Press

consulting a lexicon, we can arrive at a precise picture of which personality traits are important in a society, an approach termed the 'lexical hypothesis'. Next, researchers took large quantities of verbal descriptors of real humans and applied factor-analytic methods to analyze which trait words correlated. This was done by simply asking people whether specific words such as 'enthusiastic' or 'assertive' were something that described them. It emerged that words like 'enthusiastic' and 'assertive' tend to go hand-in-hand; a person described as 'enthusiastic' also tends to be described as 'assertive'. Super-unsupervised learning builds on the same insight. If a construct is salient, it is likely that dedicated words have been invented to describe it by the users of a language.

A word thus derives meaning from its context and the other words surrounding it. Words that tend to co-occur are more often than not similar in meaning. In essence, the approach of word embeddings works by projecting each word onto a multidimensional space, providing a word vector for each word in a corpus. The distance between words is then based on how often they tend to co-occur. If the words co-occur often, their distance is small; if the words co-occur infrequently, their distance is large. This approach can calculate the distance between different words such as 'Trump' and 'hate', the distance between a tweet (by averaging the words within it), and a construct such as 'hate'. Importantly, these distances are based on the document authors' (for example, the person who is tweeting) own perceptions of what it is related to; for example, hate and hostility. Furthermore, since this technique does not require labels, it is completely unsupervised. Accordingly, it has the same (often useful) properties that topic modelling has: labelling is not done by a few expert raters – the constructs are defined by the respondents.

Collectively, these features make word embeddings immensely useful for measuring political-hostility-in-use in a manner that complies with the three standards presented above. Word embeddings (1) automatically generate labels to measure complex, externally defined constructs at a low cost that (2) reflect the users' understanding. To derive all political texts in a corpus, a researcher would simply use the word vector for the word 'political' and calculate the distance from this vector to a text. The closer the text is to the 'political' vector, the greater the likelihood that the text is a political text. Further, we can also straightforwardly use word vectors to (3) incorporate contextual information, our third standard for measurement of online hostility. A simple way to do this is to train word embeddings separately for the contexts (for example, for Democrats and Republicans separately) before aligning the vector spaces. We elaborate on this in Appendix 1.

These three standards concern the measurement of political hostility. But the super-unsupervised approach also has several advantages when it comes to the definition and conceptualization of political hostility. Scholars fiercely disagree over how, exactly, to define political hostility: is it the use of offensive and aggressive language in online discussions; that is, 'hate speech' aimed at marginalized and vulnerable groups (Waldron 2012) or cyber-bullying (Griezel et al. 2012)? Surely the categories used by Davidson et al. (2017) – that is, 'Hate speech', 'Offensive language', and 'Neither' and the categories 'Clean' and 'Abusive' used by Djuric et al. (2015) measure quite distinct constructs. Our solution to this very difficult but important problem is simple: rely on the user's perceptions of what constitutes online hostility. If the users find that a text is hostile (that is, it has a low distance to our hostility vector), we simply take their verdict at face value. As is clear, this also implies that the SU approach is only applicable for research questions where it is relevant to accept this premise of focusing on hostility-in-use. For example, the SU approach is likely not applicable for researchers interested in quantifying violations of strict legal definitions of 'hate speech'.

The simple nature of our approach may concern some. A sceptical reader may interject: 'Are you *actually* measuring politically hostile rhetoric, or are you simply measuring whether a social media user is blaming *others* of being hostile?' An example of the latter could be a tweet saying 'Democrats are so hateful', which is not measuring hatred, whereas the tweet 'I hate all

https://doi.org/10.1017/S0007123423000042 Published online by Cambridge University Press

Democrats' would be an example of the former. However, this would be a misunderstanding of how the word vectors are created. If a tweet says 'Democrats are hateful', then the algorithm connects 'hate' and 'Democrats' since it is continuously trying to predict a word given its context words (cf. Appendix 1.1 in Appendix 1). Another tweet may say 'Neo-Nazis are racist, homophobic and filled with hate', which connects 'hate' to 'homophobic' and 'racists'. If there are more examples of the former types of tweets, 'talking about hate', rather than the latter, 'being hateful', the hate vector would not in general be associated with words that are (almost by definition) hateful, such as 'racist' and 'homophobia'.[3] The tweet 'Democrats are racist and hateful' would be a mix of being hateful, that is calling another group 'racists', and talking about hate at the same time.[4] If more tweets connect 'hate' to actual hate or simply a mix of actual hate and talk about hate, compared to tweets that are only talking about hate, the algorithm would (correctly) associate 'hate' with actual hate and not (primarily) talk about hate. Of course, we also subject the method to validity tests below to investigate whether the method is associated with actual hate but it is important to stress that the method is, by itself, likely to measure actual hate rather than talk about hate based on the simple 'distributional hypothesis': 'You shall know the word hate from the company it keeps.'[5]

Super-Unsupervised Learning and Existing Approaches

Now that we have outlined the super-unsupervised method and outlined criteria for why and when it is needed, it is worth briefly discussing similarities and differences from current approaches towards measuring constructs using quantitative text. The super-unsupervised approach uses word embeddings. Word embeddings are also used in the interesting approach 'corpus-based dictionaries', developed by Rice and Zorn (2021). Simplifying somewhat, the corpus-based dictionary (CBD) approach starts with a dictionary of an existing construct and then uses word embeddings to find words in a new set of texts that are similar to those used in the original dictionary. Both the SU approach and the CBD approach thus use word embeddings but they also differ in important aspects. In the SU approach, the measurement and definition of a construct are not defined in advance, as in the CBD approach. In the terminology used above, the CBD is a supervised technique, although with a clever twist, with all the benefits and limitations this entails.

The SU approach is also similar to the interesting Latent Semantic Scaling (LSS) approach developed by Watanabe (2021). However, there are both important conceptual and methodological differences between the SU approach and the LSS approach.[6] First of all, the LSS approach is still, in our terminology, a supervised technique since the researcher starts with a definition of their construct which the LSS can be used to measure. In the SU approach, the respondents themselves define the content of the construct. An illustrative example is 'vaccines', as we outline in Appendix 4. Before 2020 and the onset of the coronavirus pandemic, vaccines were mostly discussed in terms of classical infectious diseases such as polio and measles and were fairly apolitical. After 2019, this changed with the advent of the coronavirus pandemic. The SU approach can easily handle the fact that the conceptualization and content of constructs change over time such as in the case of vaccines, and can quantify this change, but we cannot see how this can be done in a supervised setting such as the LSS approach. Another way to express this is by the research questions the LSS and the SU approach ask. The LSS asks, 'I want to measure construct X'. The SU approach asks, 'I want to know how users define construct X'.

There are also important methodological differences between the SU and LSS approaches. The LSS approach uses a unidimensional scale to measure constructs. As a consequence, that

[3] We show below that the word vector 'hate' is, in fact, highly associated with the words 'racist' and 'homophobia'.

[4] Example tweet #5 in Fig. 2 has exactly such a sentence construction.

[5] For further discussion of this point, see Appendix 2.

[6] Watanabe (2021) also uses the phrase 'word embedding' to describe his approach, but this phrasing differs from how word embeddings are typically understood, so we prefer to refer to the method as an example of using principal component analysis on word counts; see, for example, Hobson, Howard, and Hapke (2017, chapter 4).

https://doi.org/10.1017/S0007123423000042 Published online by Cambridge University Press

approach cannot handle more complex constructs such as 'political hostility'. In contrast, the SU approach uses the power of word2vec math to combine two different dimensions into one that can capture both of these aspects, as demonstrated below.

Finally, the LSS approach cannot incorporate contextual information in the same way the SU approach can. Since the LSS approach uses principal component analysis and an additional scaling layer it cannot readily be used to compare distances across groups or over time, but this is possible with the SU approach after introducing an additional alignment step. We elaborate on the details in Appendix 1 but an example can illustrate what is at stake. For example, if we train word embedding models separately for Democrats and Republicans and we wish to see whether 'fear-mongering' is closest to the Democrat or Republican perceptions of 'political hostility', they would have to use the same scale. If not, we cannot compare this word or its relationships with any number of words across groups since we would not know the scaling across groups. Perhaps the distance is almost the same across groups or, perhaps, the distance is twice as big; using the LSS approach, there is no way of knowing.

Summing up, the LSS approach falls short of all three criteria outlined above for the measurement of 'political hostility'. It cannot be used to measure psychologically complex phenomena; it does not take its point of departure in the respondents themselves; and it cannot easily handle contextual information. Although the LSS is, we would argue, not ideal for the measurement of political-hostility-in-use, it is certainly an interesting and promising approach for other constructs, as Watanabe (2021) convincingly demonstrates.

Validating the Super-Unsupervised Approach for Studying Online Political Hostility

Our goal is to demonstrate the usefulness of the super-unsupervised approach for measuring online hostility. A crucial first step in this regard is to validate it. To this end, we use a large corpus of text data from Twitter (tweets) to benchmark our approach against multiple standard validation criteria, an approach named construct validation (Cronbach and Meehl 1955). We show that the super-unsupervised approach can be used to measure and distinguish between (1) political tweets, (2) hateful tweets, and (3) politically hateful tweets (that is the combination of the two). To our knowledge, this is the first study that demonstrates how to measure (3).

We examine face validity by extracting several concrete tweets classified as more and less politically hateful by the super-unsupervised approach to examine if it appears to measure what we claim it does. Next, we examine the convergent and discriminant validity by correlating the classifications produced by the super-unsupervised approach with several theoretically derived individual-level predictors of political hate. We examine criterion validity by extracting political actors in tweets. Then, we examine the external validity of the classifier by applying it to another dataset of Twitter posts and by directly comparing the outputs of the classifier to hand-coded labels of offensive posts on Twitter. Finally, we examine the ecological validity of the classifications by examining whether Republicans and Democrats perceive political hate differently. In particular, based on psychological research, there are reasons to believe that the objects of political animosity are different for Democrats and Republicans (Brandt et al. 2014). If the measured constructs produced by the super-unsupervised approach work as intended, such differences should be evident. As discussed above, a key feature of the super-unsupervised approach is that it allows for the integration of contextual information. The full exposition of the last four validation exercises can be found in Appendix 1.

Data and Methodology

To examine the construct validity of the super-unsupervised approach, we rely on a unique dataset of American Twitter users. These users were recruited by YouGov from their web panel and invited to participate in a study. The participants completed several surveys containing questions

https://doi.org/10.1017/S0007123423000042 Published online by Cambridge University Press

measuring their political attitudes (for example partisanship, political interest, and knowledge), self-reported online hostility, and a range of demographic variables that we analyze below. Furthermore, the respondents provided informed consent to let us scrape their Twitter data which provided us with a unique opportunity to link survey data with social media data. The Twitter scraping began on 4 March 2019 and ran until 30 March 2020, providing us with 5,076,851 tweets.

We use the tweets to create two word vectors, a 'political' word vector and a 'hate' word vector. While our object of interest is 'online political hostility', we prefer the word 'hate' over the word 'hostility' since the former word is used more often in everyday language, which our method relies on, when compared to the more technical term 'hostility'.[7] Further, a key reason for labelling our measure 'online political hostility' rather than 'online political hate' is to not conflate it with the more specific term 'hate speech', which we do not purport to measure.

In the analysis, we also use the ability to calculate a combined 'political hate' word vector by adding our two individual word vectors. Third, for each social media post in our corpus, we average the word vectors in each text and calculate the distance from each text to our three chosen word vectors: that is, 'political', 'hate', and 'political hate'. These distance measures comprise our primary measures of political hostility. If a given post is proximate to the 'hate' word vector, we thus classify it as more hateful on average. The typical distance measure is the cosine similarity measure, see also Appendix 1.1, which ranges from −1 (complete opposites) to 1 (complete similarity), with a cosine similarity of zero indicating no relationship between two vectors.

A more detailed discussion of word embeddings and our implementation can be found in Appendix 1.1. Here, we show that the super-unsupervised approach is very robust to different preprocessing choices. In addition, in Appendix 1 most of the results sections have additional information on the methods and measures, using the same headings and structure as in the main text.

Results

This section discusses the results of our investigation concerning the construct validity of the proposed super-unsupervised approach. We proceed in three steps. First, we provide simple illustrations of the face validity of the approach. Second, we provide formal tests of the convergent and discriminant validity of the outputs of the approach. Third, we provide several tests of the criterion validity of the approach. The last validity tests can be found in Appendix 1.

Face Validity

Methods and Measures

We first compare the results from our 'political' word vector and our 'hate' word vector to Google's Perspective API (JIGSAW). The Perspective API is a machine-learning classifier trained on Wikipedia editorial discussions and is a commonly applied method for detecting online toxicity (for example, Kim et al. 2021). Accordingly, comparing results using our supervised-unsupervised approach to the 'toxicity scores' obtained from applying the Perspective API on the same set of tweets provides a straightforward way of examining whether we measure what we claim to be measuring.

Results

Table 2 provides a set of example tweets with the highest scores on either toxicity or political hate (the remaining tweets are displayed in Table 4 in Appendix 1.2.2). The face validity test simply

[7] As discussed below, and illustrated in Fig. 2 in Appendix 3, the method is quite robust to the specific choice of words.

https://doi.org/10.1017/S0007123423000042 Published online by Cambridge University Press

Table 2. Full text for example tweets in Fig. 2

Number	Tweet
8	Republicans and Democrats must stand together in opposition to this sham attack against our President.
11	The Democrats have an opportunity to work with the President to bridge the divide and serve the country in the best possible way. They can't. The media is using anonymous sources to create panic.
12	It is the most savage condemnation of boomers and their profoundly capitalist ideals, warped obsession with criminal justice, irrational disdain for education, and a general false sense of responsibility I have ever seen and I cannot think of a better movie.
13	I don't like maps like this because they pretend that people like me, who hate both corrupt political parties, and the country that only works for the 1% don't exist.
23	Sen. Feinstein. I call bullshit. You've been a corporate whore for 30 years and what 'expertise' you claim has done nothing for future generations. We see you. The racist GOP needs to go as badly as the neolibs. It's time for real change.
27	The person is @IlhanMN. He is to blame for a bunch of people shooting themselves up with dirty needles. You are an idiot.
33	A big fat racist idiot doesn't stop being a big fat racist idiot because he gets lung cancer.
36	Is it appropriate to call you a white bitch? BITCH STFU.

concerns whether the tweets themselves are 'face valid'. We randomly sampled 10 example tweets from the 40,000 highest-scoring tweets for each measure. We then plotted them in a two-dimensional space according to their score on the 'political' word vector and the 'hate' word vector in Fig. 2. The third dimension, indicated by colour in the figure, shows the tweets' toxicity scores. The scales run from zero to one, where a higher score indicates a greater correspondence with the given dimension. Distributions for all variables can be seen in Appendix 1.2.

As we can see, the tweets with high toxicity scores (which are purple) are mostly about swear words such as 'idiot,' 'bitch', and 'racist'; see for example tweets 33 and 36. There are some examples of overlap. Tweet 12, which is relatively high on both the 'political' dimension and the 'hate' dimension also scores high on toxicity. This makes sense since the content is both political – it is about 'capitalists' and 'criminal justice' – and is also strongly worded but not to the extent that it also gets a high score on toxicity.

In terms of distinguishing between 'political' and 'hate', we see that tweets 8 and 11 are relatively close to the 'political' word vector, whereas they score below 0.5 on the 'hate' word vector. This also corresponds to their content: they are mostly political and not extremely hate-filled. Tweet 11, for example, talks about 'democrats' and 'the President'.

We are thus clearly picking up on something substantively interesting that goes beyond a simple algorithm that contains a dictionary lookup of, for example, the words 'hate' and/or 'political'. The distributions for toxicity and political hate can be found in Appendix 1, Fig. 1.

Table 3 shows the words most closely associated with our words vectors 'political', 'hate', and 'political hate.' It further validates the super-unsupervised approach and shows why it works better than a simple dictionary-based approach. All of the words closest to the word vectors are words we normally associate with these words. 'Hate' is closely related to, for example, 'hateful', 'racist', and 'intolerance', while 'political' is closely related to 'influence', 'opponent', and 'partisan'. Therefore, in a sense, we are not only using the words 'hate' and 'political', but we also use all the words close to them. This explains why we obtain substantively meaningful example tweets in Table 2, even when the words 'hate' or 'political' do not consistently appear. The approach is thus relatively insensitive to the specific words chosen: we could just as easily have chosen the words 'politics' or 'hostile' and we would have obtained results that are very similar to those obtained in the analyses of correlations. 'Political hate' and 'Political hostility' has a cosine similarity of 0.75.

Since we also use subwords, the results are even more robust; for each word we also include a series of subwords that include different spellings and potentially shorter forms of the same word.

https://doi.org/10.1017/S0007123423000042 Published online by Cambridge University Press

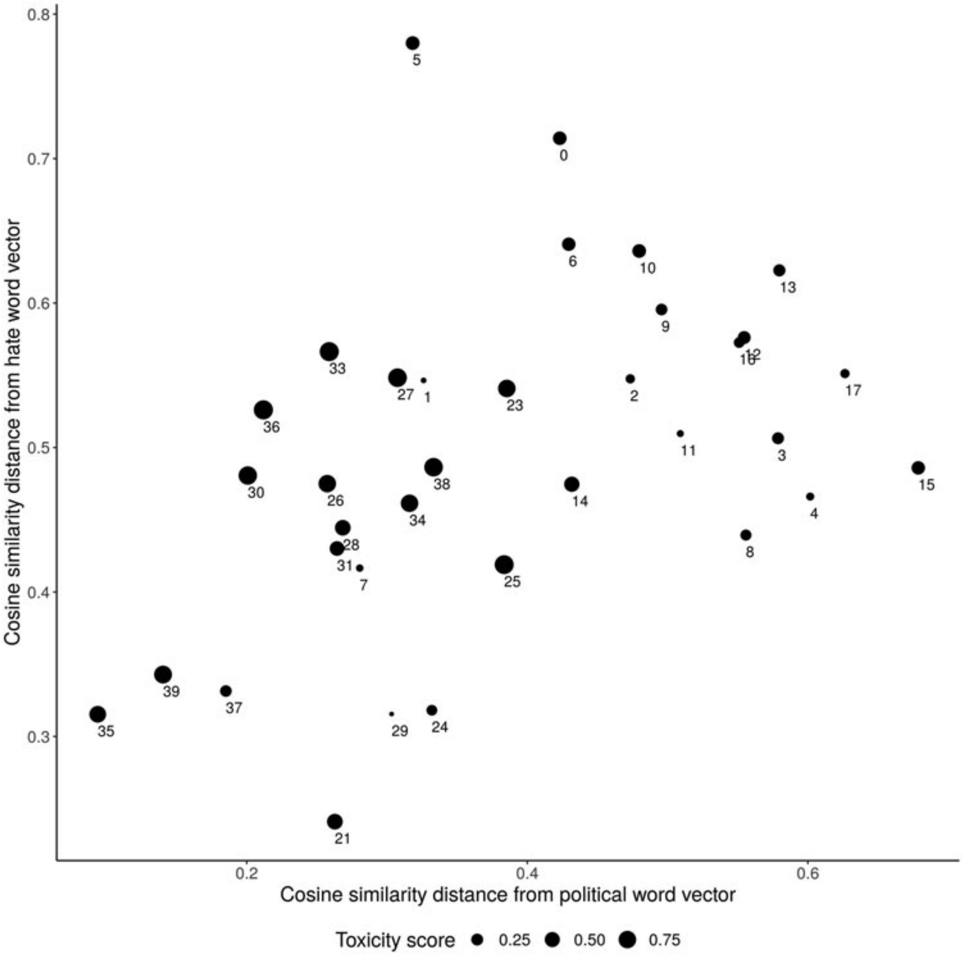

Figure 2. Sample locations for tweets in terms of their distances from the 'political' and 'hate' word vectors and their Toxicity score. The sample texts are 10 randomly selected tweets from among the first 40,000 tweets scoring closest to 'political hate' and 10 randomly selected tweets from among the 40,000 scoring most highly on toxicity. The tweets are sorted separately for toxicity and political hate. The rest of the example tweets can be found in Appendix 1.2.2.

This also applies to each of the words in the robustness check where we move further and further away from the chosen words. Taken together, all of these results suggest that our supervised-unsupervised approach is face valid.

Convergent and Discriminant Validity

Convergent and discriminant validity concerns whether our political hate vector correlates with what we expect it to correlate with (convergent validity) and whether it correlates with anything it should not (discriminant validity) (Campbell and Fiske 1959). This analysis progresses in two parts by constructing a series of marginal correlation tables. We expect the 'political hate' vector to positively correlate with several subjective, self-reported measures from our survey; in particular, political interest and political knowledge, since politically interested and knowledgeable individuals are more likely to tweet about politics (convergent validity). In addition, the 'political hate' vector should correlate positively with self-reported hostility and traditional measures of online hostility such as sentiment scores, toxicity scores, as well as more recent types of language

https://doi.org/10.1017/S0007123423000042 Published online by Cambridge University Press

Table 3. Cosine similarity distances for the word vectors 'political,' 'hate' and 'political hate.' For each word vector, the 20 words closest to the given word are extracted

Political	Hate	Political hate
(1.0, political)	(1.0, hate)	(0.83, political)
(0.68, correctness)	(0.78, hatred)	(0.79, hate)
(0.68, opponent)	(0.73, hateful)	(0.71, correctness)
(0.68, politic)	(0.66, incite)	(0.7, hatred)
(0.66, rival)	(0.66, haters)	(0.69, espouse)
(0.64, partisan)	(0.66, bigotry)	(0.69, politic)
(0.63, apolitical)	(0.64, intolerance)	(0.69, divisiveness)
(0.63, ideological)	(0.63, supremacists)	(0.66, hateful)
(0.62, politically)	(0.63, hater)	(0.65, incite)
(0.6, influence)	(0.62, supremacist)	(0.65, ideology)
(0.59, weaponize)	(0.62, homophobia)	(0.65, divisive)
(0.58, cynical)	(0.62, racists)	(0.63, weaponize)
(0.58, solely)	(0.62, foment)	(0.63, discourse)
(0.58, partisanship)	(0.61, stoke)	(0.63, ideological)
(0.57, weaponizing)	(0.61, racist)	(0.63, animosity)
(0.56, weaponized)	(0.61, incitement)	(0.62, overt)
(0.56, cartoons)	(0.61, espouse)	(0.62, intolerance)
(0.56, legitimate)	(0.61, spew)	(0.62, rhetoric)
(0.56, deference)	(0.6, racism)	(0.62, promote)
(0.55, discourse)	(0.6, supremacy)	(0.61, foment)

models (see below). At the same time, these traditional measures should be less related to the survey-based measures of political interest and political knowledge when compared to our 'political hate' vector (that is discriminant validity) as these traditional measures are not oriented towards politics specifically. We also include demographic information on gender since this is often discussed in terms of online hostility (Siegel 2020).

To compare the super-unsupervised approach with a traditional classification approach we hand-label 2,500 tweets, which further corroborates the convergent validity of the super-unsupervised approach. This allows us to compare the results to the state-of-the-art approach to measuring online hate, the 'binary classification task' (Siegel 2020, 59). We use hand-labelling as a first step and add a classifier as a second step.

In the second part of the analysis, we calculate a series of partial correlations to investigate the two separate dimensions of political hate, namely 'political' and 'hate'. After partialing out the shared variance between the 'political' vector and the 'hate' vector, the correlation between the 'political' vector and hostility and toxicity should be much smaller than the marginal correlation above. Conversely, the 'hate' vector should correlate much less with political interest and political knowledge when the shared variance with the 'political' vector is accounted for.

Methods and Measures

To measure sentiment, we apply the popular AFINN sentiment dictionary to our tweets (Nielsen 2011). AFINN scores rest on a dictionary approach, giving valence scores from −5 (very negative) to +5 (very positive) to each word in the dictionary. This is by no means the only way to measure sentiment but it serves as a good benchmark to contrast our estimates (Stine 2019). We also use Google's Perspective API toxicity scores, as mentioned above, as a second external measure of online hate.

We asked four expert raters[8] to rate 2,500 randomly sampled tweets from the full corpus. Each rater received 500 common and 500 unique tweets and was asked to rate how hostile (five-point

[8] The raters were a mix of postdocs and PhD students working in the field of online hostility from INSTITUTION BLINDED but not part of this particular research paper.

https://doi.org/10.1017/S0007123423000042 Published online by Cambridge University Press

Table 4. Correlation table for the correlations between measures based on tweets (political hate, political, hate, toxicity and sentiment) and the measures from the survey data (political interest, political knowledge, self-reported hostility and gender

	Toxicity score	Afinn score	Political vector	Hate vector	Political interest	Political knowledge	Hostility	Female
Political hate expert	0.35	−0.28	0.67	0.45	0.43	0.37	0.28	−0.11
Political hate	0.47	−0.32	0.91	0.87	0.33	0.35	0.24	−0.07
	0.67	−0.61	0.94	0.9	0.41	0.44	0.3	−0.13
Political hate classifier	0.34	−0.25	0.66	0.43	0.42	0.36	0.27	−0.12
	0.57	−0.55	0.81	0.69	0.42	0.42	0.28	−0.1
Political hate context	0.39	−0.3	0.63	0.47	0.32	0.33	0.24	−0.13
	0.59	−0.62	0.84	0.65	0.43	0.44	0.27	−0.17
Toxicity score	1.0	−0.24	0.4	0.45	0.14	0.17	0.15	0.0
	1.0	−0.68	0.58	0.67	0.24	0.22	0.29	−0.05
Afinn score	−0.24	1.0	−0.3	−0.27	-0.14	−0.16	−0.12	0.07
	−0.68	1.0	−0.59	−0.53	-0.28	−0.29	−0.27	0.11
Political vector	0.4	−0.3	1.0	0.6	0.36	0.38	0.22	−0.12
	0.58	−0.59	1.0	0.7	0.45	0.48	0.27	−0.21
Hate vector	0.45	−0.27	0.6	1.0	0.21	0.23	0.22	−0.0
	0.67	−0.53	0.7	1.0	0.28	0.31	0.29	−0.01
Political interest	0.14	−0.14	0.36	0.21	1.0	0.52	0.24	−0.15
	0.24	−0.28	0.45	0.28	1.0	0.48	0.17	−0.18
Political knowledge	0.17	−0.16	0.38	0.23	0.52	1.0	0.23	−0.27
	0.22	−0.29	0.48	0.31	0.48	1.0	0.17	−0.3
Hostility	0.15	−0.12	0.22	0.22	0.24	0.23	1.0	−0.13
	0.29	−0.27	0.27	0.29	0.17	0.17	1.0	−0.12
Female	0.0	0.07	−0.12	−0.0	−0.15	−0.27	−0.13	1.0
	−0.05	0.11	−0.21	−0.01	−0.18	−0.3	−0.12	1.0

The first row for each construct, which has a grey colour, is the result of the annotation sample, and the second, which has a black colour, is the result of the full sample.

scale, 1 = 'Not at all hostile', 5 = 'Very hostile') and political (binary scale, 0 = 'Not political', 1 = 'Political') each tweet was. We aggregated the ratings to create a measure of 'political hate', labelled 'Political hate expert' in Table 4. We use this measure to create a classical predictive model using the labelled data as input, which we subsequently used to score the rest of the tweets (see SM Section, Appendix 1.3 for further details on the expert raters and the classical classification exercise). If we treat the political hate measure and the expert-based measure of political hate as binary by splitting them at the mean, we get an accuracy of 74 against a baseline of 50 since our groups are balanced, which is quite good. They are thus measuring the same construct to a very large extent and the idea that the super-unsupervised approach is merely measuring 'talking about hate' is thus ill-founded.

Finally, we include a measure of political hostility based on so-called pre-trained language models that follow a similar approach for calculating the distance from 'political hate' to each tweet to gauge whether the tweet is politically hostile or not. These newer pre-trained language models can take into account more advanced grammatical relationships. Whereas each word has one single location in a vector space in the super-unsupervised approach, a word's location in these models depends on the surrounding words. A prime example is the word 'bank' which, depending on the surrounding words, can refer either to a 'river bank' or a place to withdraw and deposit money. Because of this ability, the embeddings here are sometimes referred to as 'contextual' embeddings (Clark et al. 2019; Hewitt and Manning 2019). We therefore term this measure 'Political hate – language model' in Table 4. The most famous of these is probably Google's BERT model (Devlin et al. 2018).[9] The main drawback of this approach is the amount of data needed is

[9] We use a specific version of this architecture that is ideally suited to calculating (contextual) embeddings for sentences that we are interested in here. See Appendix 1.3.

https://doi.org/10.1017/S0007123423000042 Published online by Cambridge University Press

potentially huge, which makes it difficult to use to study changes in embeddings across, for example, Republicans and Democrats, or over time as further discussed below in Section 'Ecological validity: Context matters'.

Marginal Correlations with Self-Reported Measures and Expert Raters

Table 4 presents the results. Here, the first row for each construct gives the results for the annotation sample (grey), while the second row (black) shows the results for the full sample. It is clear that the correlations for the tweets from the annotation exercise are slightly smaller than those in the full sample; this is likely because the variation is slightly smaller. Importantly, however, the relative performance of the different classifiers remains essentially the same for the two samples.

Turning to the data from the full sample, we see that the super-unsupervised approach performs just as well as the 'political hate classifier' and the more advanced 'political hate context classifier'. The correlations between our survey-based measures and these three different operationalizations of political hate are almost identical. And the correlations between these three operationalizations of political hostility are also very high, hovering at around 0.8, suggesting they measure a very similar construct. Results for our 'political' word vector are similarly strong: The correlations between the 'political' vector on the one hand and self-reported political interest and political knowledge on the other, are 0.45 and 0.48, respectively. Even in a survey setting using only self-reported measures, a correlation of 0.48 is high. The correlations presented here are correlations for completely different ways of measuring political interest; that is, self-reported political interest as measured by a survey and interest in politics measured by how often somebody engages in political discussions on Twitter. These findings corroborate the strong convergent validity of the super-unsupervised approach.

The 'hate' word vector also correlates positively with the expected measures: the correlation with self-reported hostility is 0.30 and the correlation between political hate and gender is −0.13. A correlation for a binary variable such as gender is perhaps not the most intuitive way of representing the results, however. If we split the 'political hate' column into six equally large segments, the proportion of females is 0.64 for those who are least hateful and 0.42 for those who are most hateful. This corresponds to previous studies demonstrating that those who produce hate speech are more likely to be male (Costello and Hawdon 2018)

The results for the 'hate' word vector are fairly similar to those obtained with both the toxicity and sentiment scores; both correlate at roughly the same levels with hostility. Neither the toxicity nor sentiment scores correlate as highly with political knowledge and political interest. If we were to use these to capture political hate, we would probably be misled since the correlations for our 'political hate' word vector and political interest and political knowledge are roughly double those of the correlations that toxicity and sentiment have with these measures. Thus, although the toxicity and sentiment scores seem to capture the 'hate' component, they fail to capture the important 'political' part of online political hate. Accordingly, these results show that our supervised-unsupervised approach has strong divergent validity. Figure 1 in Appendix 3 suggests a sample size of 100,000 tweets is needed for our approach to work.

Ecological Validity: Context Matters

To illustrate the usefulness of the super-unsupervised approach and to indirectly validate it further, we investigate its ability to take context into account. This is a test of ecological validity; that is, whether investigations can capture perceptions of participants in their 'real-life' environments (Schmuckler 2001).

At least two avenues for research open up using this approach. First, we can examine expressions of political hate across different contexts at a single point in time. For example, we can examine whether Democrats and Republicans express political hate differently and whether

https://doi.org/10.1017/S0007123423000042 Published online by Cambridge University Press

political hate differs across countries. We know that Democrats and Republicans perceive the world differently at a fundamental level (Jost, Federico, and Napier 2009), so we could expect differences in their perceptions of political hate. Further, psychologists have recently used word embeddings to investigate the presence of gender discrimination across forty-five different languages (DeFranza, Mishra, and Mishra 2020), and it would also make sense to expect differences in political hate across countries. Second, we can study changes in political hate in the same context over time. An example of such an approach is Garg et al. (2018), who use word embeddings to study changes in gender and ethnic stereotypes over time.

It is important to note that we are not the first to demonstrate the usefulness of using word embeddings to study changes over time or across contexts. What we are claiming is that no currently existing method for measuring online hate can specifically do this. This is why the super-unsupervised approach is an important contribution to the measurement of online political hostility. Although in principle, expert raters in different countries could be used to train a classifier, it would be very costly and difficult to compare levels of political hate across countries. This is in contrast possible using word embeddings if the vector spaces are aligned; we elaborate on this in Appendix 1.1. In principle, it is also possible to use newer methods utilizing context embeddings to study the changes in embeddings over time, but in practice it is quite difficult since these methods require much more data than is available in most real-life use cases; a recent study suggested that 10 million to 100 million words are needed (Zhang et al. 2020). We suggest that the super-unsupervised approach represents a great alternative to these two techniques in many situations.

Methods and Measures

To illustrate the usefulness of the approach, we perform two analyses. First, we use our American Twitter data to create two classifiers for political hostility separately for participants who identify as either Republicans or Democrats. We then align their vector spaces and investigate which words are differentially related to political hostility for these two subpopulations (see Appendix 1.5 for general details and Appendix 1.6.5 and Appendix 1.6.6 for specific results).

Second, we use Twitter data collected before and after recent national elections in four countries (Denmark, Sweden, Germany, and Italy) to demonstrate the approach's ability to study changes over time and across contexts. In each country, we follow our supervised-unsupervised approach but use local words for 'political hate' to construct the word vectors (see Appendix 1.6 for details).[10] We then examine over-time changes in political hate within each country. Intuitively, if our approach captures the relevant contextual features, we would expect to find that national levels of political hate rise as the election day draws nearer, before dropping as the political dust settles.

Results

Figures 3 reveal how strikingly differently Democrats and Republicans think about 'political hate'. Much in line with lay intuitions about modern-day cleavages in American politics, we find that Democrats are more likely than Republicans to associate political hate with 'fearmongering', 'Trumpian', and 'rightwing', while Republicans, on the other hand, place emphasis on words like 'left' and 'democrat'. It is also noteworthy that 'purity', a moral foundation valued particularly by Republicans (Haidt 2013), shows up in the word 'disgust'. This is in line with studies demonstrating greater disgust sensitivity for Republicans (Inbar, Pizarro, and Bloom 2009). These different conceptions of political hate across partisan subpopulations are important. Not only do they demonstrate just how differently people of different political stripes can understand the

[10] The words used to measure 'political' across Denmark, Sweden, Germany, and Italy were 'politik', 'politik', 'politik', and 'politica'; the words used to measure 'hate' were 'had', 'hata', 'hassen', and 'odiare'.

https://doi.org/10.1017/S0007123423000042 Published online by Cambridge University Press

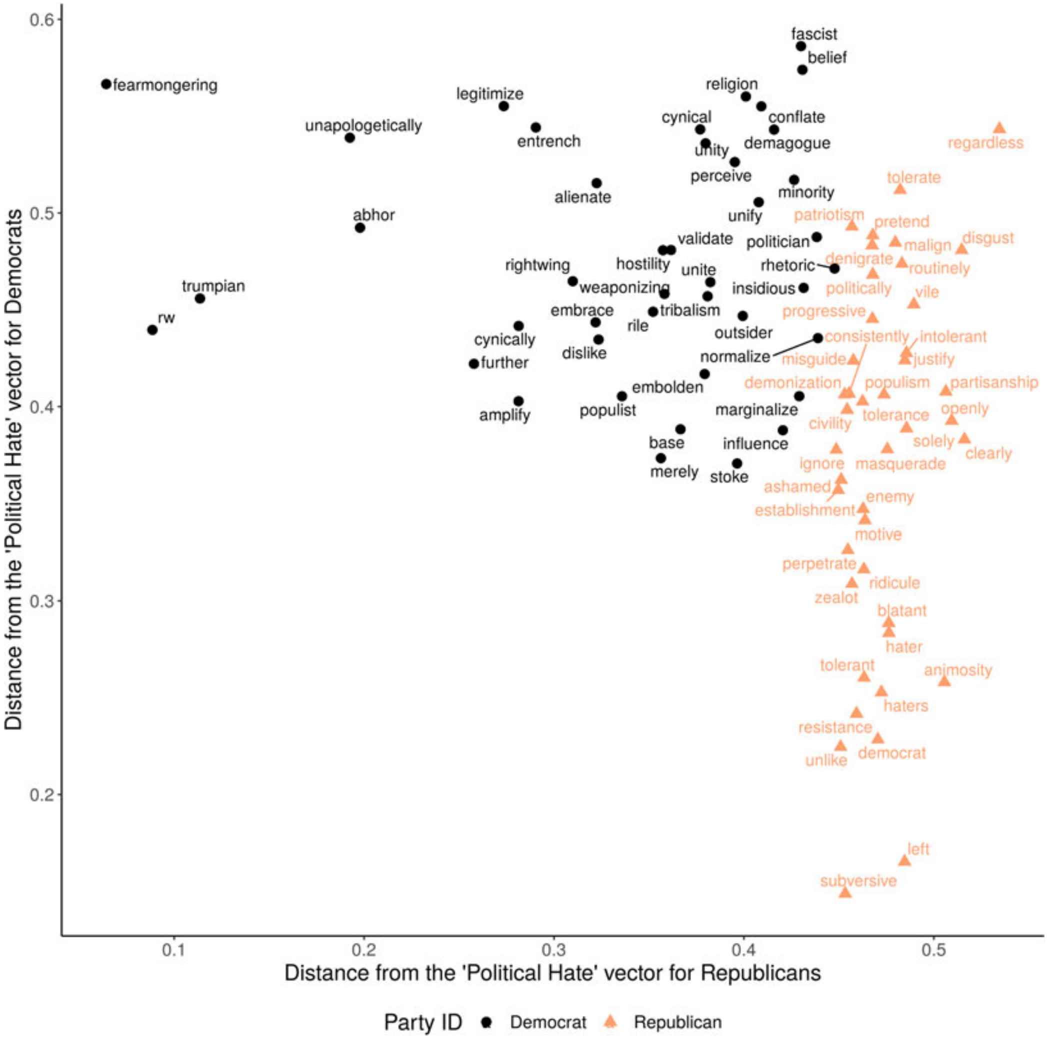

Figure 3. The figure illustrates the distance from each word to Political Hate for the Democratic and the Republican classifier of Political Hate. We have included the top 70 words for each group and excluded those that were common to each.

same political concepts but they also underscore one of our key theoretical arguments: context matters significantly for how we understand words.

Figure 4 illustrates the development of political hate leading up to general elections in Sweden, Denmark, Germany, and Italy. In all cases, we can see that political hate reaches its maximum on the exact day of the election and that political hate slowly increases during the period leading up to the election. This validity check demonstrates our super-unsupervised approach is indeed picking up on political hate. But it also illustrates the power of the approach to study changes over time and across different contexts with little effort; an objective that would be difficult to accomplish with, say, dictionary-based approaches tailored to the English language.

Additional Validity Tests

The full background and some additional results for the remaining validity tests – that is, criterion validity, external validity, and ecological validity – can be found in Appendix 1.

https://doi.org/10.1017/S0007123423000042 Published online by Cambridge University Press

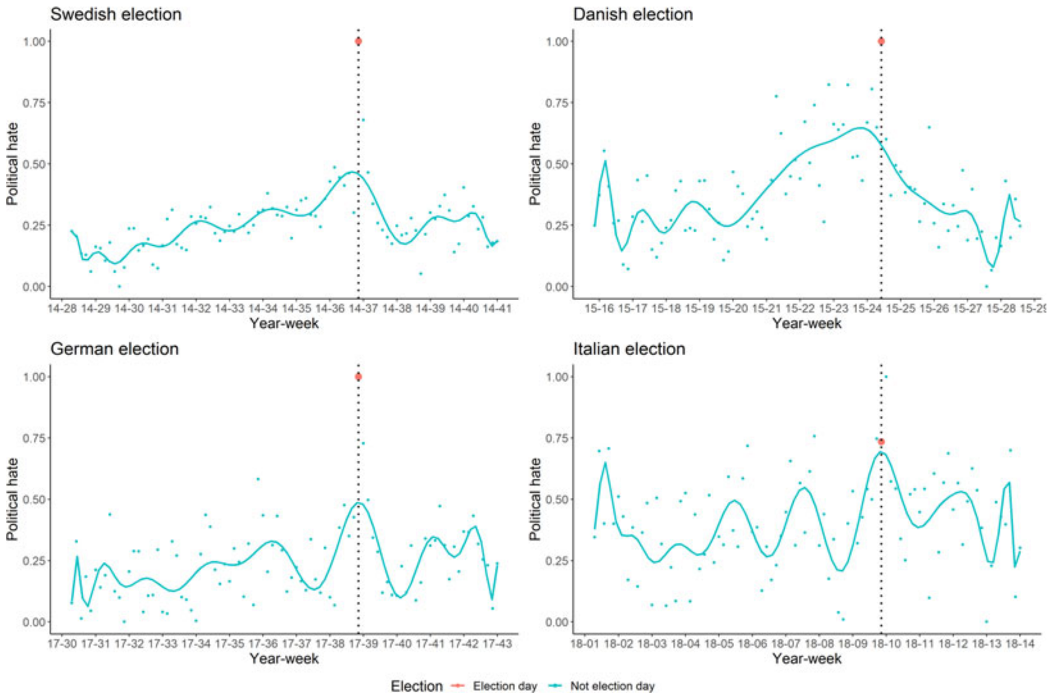

Figure 4. The figure illustrates the development of political hate on Twitter in the period leading up to general elections in Sweden, Denmark, Germany, and Italy in 2014, 2015, 2017, and 2018, respectively.

Conclusion and Discussion

Current approaches to studying online (political) hostility are struggling with measurement and definition to such an extent that a recent review of the literature has called it being 'clouded by definitional ambiguity' (Siegel 2020, 59). This has hindered a cumulative research programme on the causes and consequences of online hostility as well as efforts to fight its continued presence on social media platforms. In this manuscript, we have argued that the super-unsupervised approach to measuring and defining political hostility has clear advantages when compared to classic methods such as topic modelling approaches, dictionary-based approaches, and approaches using a classifier based on the meticulous hand-labelling of texts. We first set up three standards for the measurement of online political hostility and argue that they have clear advantages compared to current state-of-the-art approaches.

By focusing on hostility-in-use, that is users' everyday word usage in Twitter communications, we can attach the label of 'political hostility' to an otherwise completely unlabelled set of tweets. In this way, we can circumvent the high cost of hand-labelling, which haunts the supervised approach and remedy the inability of the supervised approach to handle context-dependent labels. This also carries an important implication for the current difficulties of defining and measuring hate speech in online (social) media. If the users do not have a single clear idea of what hostile speech is, trying to measure this as a single construct applicable to all contexts will always prove difficult, even when using very advanced NLP techniques. Researchers can define external 'classifications' of hostility using hand-labelling but, in a sense, they are bound to fail if these conceptualizations do not, to some extent, correspond to the way users think and write about the construct – possibly in quite different ways in different situations.

The super-unsupervised approach not only provides meaningful 'classifications' but also beats current standard methods used to measure online political hostility such as dictionary-based approaches (AFINN scores) and model-based classifications (toxicity scores). To demonstrate this, we subjected the approach to five validity tests: face validity, convergent and discriminant validity, and criterion validity; in the Appendix, we discuss the external validity and further ecological validity tests of the approach. We can distinguish between the concepts of (1) 'hate', (2) 'political', and (3) 'political hate' using this approach so, depending on the researcher's need, there are several options from which to choose. From a methodological point of view, this highlights the fact that the method can be used to study related but distinct constructs with a fair amount of precision.

Furthermore, the approach has clear advantages in terms of the definition and conceptualization of political hate: we simply use the user's perceptions of what constitutes a politically hateful text. By taking this approach, we side-step the difficult conceptual task of defining and conceptualizing what constitutes (political) hate online, which plagues dictionary-based approaches as well as approaches based on the hand-labelling of texts.

In Appendix 5 we outline all the steps needed to use the SU approach for new constructs.[11] A few comments on the general outline of the method are in order here. Most importantly, the method should only be used if researchers want to measure how users perceive a construct; not, for example, legal definitions of a construct. This is not necessarily limited to simple constructs consisting of a single word but can also consist of combinations of words such as 'political' and 'hate' as we have done here. Once the initial analytical step has been performed, we would urge researchers to conduct face validity tests where we investigate the 'nearest neighbours' (3) as well as example tweets (2). These investigations both serve as validity tests and also serve as a way of understanding how users conceptualize the construct. We discuss this in more detail and what to do when these validity tests fail in Appendix 5.

This approach, and our validation, face limitations of course. First, there is the practical limitation that the method requires a sample size of around 100,000 tweets to reliably estimate the effects found here. In some cases, it is possible to simply use pre-trained embeddings such as those available through pre-trained language models, as discussed by Rodriguez and Spirling (2022). However, there is one inherent difficulty in doing this, seen from the perspective of the SU approach: we no longer rely on the users' language-in-use and thus cannot use the SU to conceptualize constructs and can only measure pre-defined categories.[12] Although Denny and Spirling (2018, 113) suggest that 'there is little evidence [that] using pretrained embeddings is problematic for subdivisions of the corpus by party' because 'Human coders generally prefer pretrained representations', this is, in the view of the SU approach, putting the cart before the horse. Embeddings reflect 'language in use' by, for example, Twitter users or parliamentarians when making speeches whereas 'human coders', who process the data, are not part of this language domain. Whether or not a set of crowd-sourced raters 'generally prefer' one set of words over another is irrelevant. When we do break down embeddings by party or time, we see very big differences when it comes to the conceptualizations of, for example, abortion and vaccines (see Appendix 4).

A second practical limitation of the present investigation is that we have primarily validated the approach using Twitter data. Twitter has over three hundred million users and is one of the world's most important online political platforms. It is relevant to assess this approach on other social media platforms such as Facebook as well as in completely different political contexts such as parliamentary debates and even in non-political settings such as Reddit.

There are also relevant theoretical limitations relating to the SU approach. First, this method utilizes the conception that words acquire meaning in the context in which they appear and is thus an improvement over simple word counts such as dictionary-based methods (for example,

[11] And an example demonstration on how to use the method can be found on OSF.

[12] We also elaborate on using pre-trained embeddings in the context of the SU approach in Appendix 1.5 and Appendix 4.2.

https://doi.org/10.1017/S0007123423000042 Published online by Cambridge University Press

AFINN scores). However, the word2vec method still does not take word order into account (Chang and Masterson 2020) and more advanced grammatical relationships that might affect the meaning of words (Clark et al. 2019; Hewitt and Manning 2019). As discussed above, more advanced pre-trained language models such as Google's BERT exist, which did not provide better or worse results than the SU approach in terms of convergent and discriminant validity. When data is plentiful, that is in the hundreds of millions, we would, however, recommend training these models from scratch to achieve a more nuanced and fine-grained understanding of political hate. Second, as the approach is premised on a focus on language-in-use, the use of this approach is limited to the research questions where this premise is fruitful. This includes a wide range of psychological research questions related to the causes and consequences of, for example, online political hostility. In contrast, to quantify the number of strict legal violations of 'hate speech' laws, the approach cannot stand alone. However, even when research questions involve reliance on clearly-defined constructs, the SU approach may still be relevant to some extent as the approach can be used as a first step in hand-labelling. Imagine a study of the frequency of hate speech using a strict legal definition. The use of respondents' perceptions via the SU approach would not be sufficient but it could serve as a powerful first screening step, which could provide expert coders with a sample to code. Instead of having to go through millions and millions of tweets, this could serve as a first rough attempt at separating tweets into hateful and non-hateful. This could be combined with, for example, the active-labelling approach of Miller, Linder, and Mebane (2020).

In general, we suggest that the SU approach can be used to measure any construct that is also used in natural language and is distinct enough. This is especially true when the data are plentiful – that is in the millions – such as is often the case with online social media data. Here, the abundance of data is both an advantage for the SU approach and an obstacle to supervised approaches as the sheer size will limit the possibilities of hand-labelling. Future studies should thus seek to validate the approach for other constructs in addition to 'political hostility'.

Supplementary material. Online appendices are available at: https://10.1017/S0007123423000042.

Data Availability Statement. Replication Data for this article can be found in Harvard Dataverse at https://doi.org/10.7910/DVN/4X7IMW.

Acknowledgements. We would like to thank Robert Klemmensen and Matt Loftis for their helpful comments in writing this article. In addition, we would like to thank the three anonymous reviewers for their help in writing this article; especially the first reviewer who provided extremely helpful and detailed suggestions, which improved the paper a lot.

Financial Support. This study was funded by grant AUFF-E-2016-9-22 from the Aarhus University Research Foundation to Michael Bang Petersen and grant CF18-1108 from the Carlsberg Foundation to Michael Bang Petersen.

Conflicts of Interest. None.

References

Barberá P et al. (2021) Automated text classification of news articles: A practical guide. *Political Analysis* 29(1), 1–24.

Blei DM, Ng AY and Jordan MI (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research* **3**(Jan), 993–1022.

Bor A and Petersen MB (2022) The Psychology of Online Political Hostility: A Comprehensive, Cross-National Test of the Mismatch Hypothesis American political science review, 116(1), 1–18.

Brandt MJ et al. (2014) The ideological-conflict hypothesis: Intolerance among both liberals and conservatives. *Current Directions in Psychological Science* **23**(1), 27–34.

Campbell DT and Fiske DW (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* **56**(2), 81.

Chang C and Masterson M (2020) Using word order in political text classification with long short-term memory models. *Political Analysis* **28**(3), 395–411.

Clark K et al. (2019) What Does Bert Look at? An Analysis of Bert's Attention. *arXiv preprint arXiv:1906.04341.*

Costello M and Hawdon J (2018) Who are the online extremists among us? Sociodemographic characteristics, social networking, and online experiences of those who produce online hate materials. *Violence and Gender* **5**(1), 55–60.

Cranmer SJ and Desmarais BA (2017) What can we learn from predictive modeling? *Political Analysis* **25**(2), 145–166. https://doi.org/10.1017/pan.2017.3.

Cronbach LJ and Meehl PE (1955) Construct validity in psychological tests. *Psychological Bulletin* **52**(4), 281.

Davidson T et al. (2017) Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th international AAAI conference on web and social media*, 512–15. ICWSM '17. Montreal, Canada.

DeFranza D, Mishra H and Mishra A (2020) How language shapes prejudice against women: An examination across 45 world languages. *Journal of Personality and Social Psychology* **119**(1), 1–7.

Denny MJ and Spirling A (2018) Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis* **26**(2), 168–89.

De Saussure F (2011[1916]) *Course in General Linguistics*. Columbia: Columbia University Press.

Devlin J et al. (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv 1810.04805*.

Djuric N et al. (2015) Hate Speech Detection with Comment Embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, 29–30.

Duarte JL et al. (2015) Political diversity will improve social psychological science 1. *Behavioral and Brain Sciences* **38**, 1–58.

Firth J (1957) *Studies in Linguistic Analysis*. Oxford: Oxford University Press.

Garg N et al. (2018) Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* **115**(16), E3635–E3644.

Griezel L et al. (2012) Uncovering the structure of and gender and developmental differences in cyber bullying. *The Journal of Educational Research* **105**(6), 442–55.

Grimmer J (2010) A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis* **18**(1), 1–35.

Haidt J (2013) *The Righteous Mind: Why Good People are Divided by Politics and Religion*. New York: Pantheon Books.

Hewitt J and Manning CD (2019) A Structural Probe for Finding Syntax Inword Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies, volume 1 (long and short papers)*, 4129–38.

Hobson L, Howard C and Hapke HM (2017) *Natural language processing in action*, New York: Manning.

Inbar Y, Pizarro DA and Bloom P (2009) Conservatives are more easily disgusted than liberals. *Cognition and Emotion* **23** (4), 714–25.

JIGSAW. Available from https://perspectiveapi.com/#/home.

Jost JT, Federico CM and Napier JL (2009) Political ideology: Its structure, functions, and elective affinities. *Annual Review of Psychology* **60**, 307–37.

Kim JW et al. (2021) The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication* **71**(6), 922–46.

Matthews G, Deary IJ and Whiteman MC (2003) *Personality Traits*. Cambridge: Cambridge University Press.

Meredith W (1993) Measurement invariance, factor analysis and factorial invariance. *Psychometrika* **58**(4), 525–43.

Miller B, Linder F and Mebane WR (2020) Active learning approaches for labelling text: Review and assessment of the performance of active learning approaches. *Political Analysis* **28**(4), 532–51.

Muddiman A, McGregor SC and Stroud NJ (2019) (re) claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries. *Political Communication* **36**(2), 214–26.

Nielsen FÅ (2011) Anew Anew: Evaluation of Award List for Sentiment Analysis in Microblogs. *arXiv preprint arXiv:1103.2903*.

Rasmussen J (2022) The Hate Speech Consensus: Severity Shapes Americans' and Danes' Preferences for Restricting Hate Speech, Psyarxiv, https://doi.org/10.31234/osf.io/j4nuc.

Rasmussen SHR et al. (2023) Replication Data for: "Super-Unsupervised" Classification for Labeling Text: Online Political Hostility as an Illustration. Available from https://doi.org/10.7910/DVN/4X7IMW, Harvard Dataverse, V1.

Rheault L and Cochrane C (2020) Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis* **28**(1), 112–33.

Rice DR and Zorn C (2021) Corpus-based dictionaries for sentiment analysis of specialized vocabularies [in en]. *Political Science Research and Methods* **9**(1), 20–35. ISSN: 2049–8470, 2049–8489. https://doi.org/10.1017/psrm.2019.10.

Roberts ME et al. (2013) The Structural Topic Model and Applied Social Science. In *Advances in neural information processing systems workshop on topic models: Computation, application, and evaluation*, vol. 4. Harrahs and Harveys, Lake Tahoe.

Rodman E (2020) A timely intervention: Tracking the changing meanings of political concepts with word vectors. *Political Analysis* **28**(1), 87–111.

Rodriguez PL and Spirling A (2022) Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics* **84**(1), 101–15. ISSN: 0022-3816. https://doi.org/10.1086/715162.

https://doi.org/10.1017/S0007123423000042 Published online by Cambridge University Press

Rodriguez A, Argueta C and Chen Y-L (2019) Automatic Detection of Hate Speech on Facebook Using Sentiment and Emotion Analysis. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 169–74. IEEE.

Schmuckler MA (2001) What is ecological validity? A dimensional analysis. *Infancy* **2**(4), 419–36.

Siegel AA (2020) Online hate speech. In Persily N and Tucker JA (eds), *Social media and Democracy: The State of the Field, Prospects for Reform*. Cambridge: Cambridge University Press, pp. 56–88.

Siegel AA, et al. (2021) Trumping hate on twitter? Online hate speech in the 2016 U.S. election campaign and its aftermath. *Quarterly Journal of Political Science* **16**(1), 71–104.

Stine RA (2019) Sentiment analysis. *Annual Review of Statistics and its Application* **6**, 287–308.

Theocharis Y et al. (2020) The dynamics of political incivility on Twitter [in en]. *SAGE Open* **10**(2), 2158244020919447. ISSN: 2158–2440. https://doi.org/10.1177/2158244020919447.

Törnberg A and Törnberg P (2016) Muslims in social media discourse: Combining topic modeling and critical discourse analysis. *Discourse. Context & Media* **13**, 132–42.

Waldron J (2012) *The Harm in Hate Speech*. Harvard: Harvard University Press.

Watanabe K (2021) Latent semantic scaling: A semisupervised text analysis technique for new domains and languages. *Communication Methods and Measures* **15**(2), 81–102. ISSN: 1931–2458. https://doi.org/10.1080/19312458.2020.1832976.

Wich M, Bauer J and Groh G (2020) Impact of Politically Biased Data on Hate Speech Classification. In Proceedings of the fourth workshop on online abuse and harms, 54–64. Online: Association for Computational Linguistics, November. https://doi.org/10.18653/v1/2020.alw-1.7. Available from https://aclanthology.org/2020.alw-1.7.

Zhang Y et al. (2020) When do you Need Billions of Words of Pretraining Data? *arXiv preprint arXiv*:2011.04946.

Cite this article: Hebbelstrup Rye Rasmussen S, Bor A, Osmundsen M, Petersen MB (2024). 'Super-Unsupervised' Classification for Labelling Text: Online Political Hostility as an Illustration. *British Journal of Political Science* **54**, 179–200. https://doi.org/10.1017/S0007123423000042

https://doi.org/10.1017/S0007123423000042 Published online by Cambridge University Press