Style over substance: A psychologically informed approach to feature selection and generalisability for author classification

Isabel Holmes, Timothy Cribbin, Nelli Ferenczi

PII: S2451-9588(22)00101-4

DOI: https://doi.org/10.1016/j.chbr.2022.100267

Reference: CHBR 100267

To appear in: *Computers in Human Behavior Reports*

Received Date: 25 October 2021

Revised Date: 14 December 2022

Accepted Date: 20 December 2022

Please cite this article as: Holmes I., Cribbin T. & Ferenczi N., Style over substance: A psychologically informed approach to feature selection and generalisability for author classification, *Computers in Human Behavior Reports* (2023), doi: https://doi.org/10.1016/j.chbr.2022.100267.

Style over substance: A psychologically informed approach to feature selection and generalisability for author classification

Isabel Holmes[1], Timothy Cribbin[1], Nelli Ferenczi[2]

[1] Department of Computer Science, Brunel University London

[2] Department of Psychology, Brunel University London

Style over substance: A psychologically informed approach to feature selection and generalisability for author classification

Abstract

Author profiling, or classifying user generated content based on demographic or other personal attributes, is a key task in social media-based research. Whilst high-accuracy has been achieved on many attributes, most studies tend to train and test models on a single domain only, ignoring cross-domain performance and research shows that models often transfer poorly into new domains as they tend to depend heavily on topic-specific (i.e., lexical) features. Knowledge specific to the field (e.g., Psychology, Political Science) is often ignored, with a reliance on data driven algorithms for feature development and selection.

Focusing on political affiliation, we evaluate an approach that selects stylistic features according to known psychological correlates (personality traits) of this attribute. Training data was collected from Reddit posts made by regular users of the political subreddits of r/republican and r/democrat. A second, non-political dataset, was created by collecting posts by the same users but in different subreddits.

Our results show that introducing domain specific knowledge in the form of psychologically informed stylistic features resulted in better out of training domain performance than lexical or more commonly used stylistic features.

Keywords

author profiling, political affiliation classification, stylistic feature sets, model generalisability, political psychology, feature development, interdisciplinarity, domain-specific knowledge

1. Introduction

Researchers are increasingly interested in what we can discover about a person from their writing. What can a person's posts on social media, for example, reveal about their social group, attitudes or personality? For instance, can we group individuals by gender purely based on their blog posts? These questions fall under the heading of author profiling, a sub task of author analysis, which involves inferring demographic and other personal attributes about the authors of a text. This area has become increasingly diverse in terms of target attributes, with studies now covering a range of domains including political science and psychology (Hinds and Joinson, 2018; Oberlander and Gill,

2004; Yu et al., 2008). In particular, the topic of political affiliation classification has been addressed many times. Here the task is to label an author (or speaker if speeches are used) by their political affiliation or outlook. In the United States, for example, it might be a binary task - Republican or Democrat – although in some circumstances potential political affiliations or outlooks may sometimes involve a higher cardinality i.e., a multiclass task (Gu and Jiang, 2021; Yu and Diermeier, 2010). The task of inferring political affiliation typically involves the use of machine learning algorithms which must be trained using features extracted from text.

Developing a good feature-set is key for ensuring a model performs well, as summarised by the axiom “garbage in, garbage out”. Approaches in political classification have been varied. For example, researchers have made wide use of ‘bag-of words’ methods such as TFIDF, a way of weighting the importance of words by their frequency, when vectorising text and selecting which words to use as inputs (Yu et al., 2008; Yu and Diermeier, 2010). Vector representations comprised of literal word or unigram counts can in some cases make up the entirety of the feature-set. More recent work has utilised more sophisticated word-embedding approaches such as GloVE (Pennington et al., 2014) and also stylistic features and non-textual features such as retweeting or mentioning (Das et al., 2021). Typically, it is model performance rather than a priori hypotheses (i.e., a data-driven approach) that is used to determine which features are likely to be most effective at discriminating the classes. However, our research shows that little to no attention has been paid to what attribute domain-specific knowledge could do to benefit data science work in this area.

Several traits are known to correlate with conservative or liberal beliefs, and previous work has found that traits such as grandiose narcissism manifest in writing (Cutler et al., 2021; Jost et al., 2003; Kruglanski, 1996; Zavala et al., 2010). We therefore argue here that more valid models might result if we use this knowledge of predictive traits and their likely expressive manifestations to inform the specification of features.

A second, related problem in political inference modelling, we argue, is that model performance is usually assessed only on text from a very similar topical domain to that used training. This means it is

often difficult to know how well a model will generalise to new observations or, indeed, if is actually measuring political affiliation rather than some other trait or attitude. In this paper, our experimental results show how model performance within a training domain of political discourse is not necessarily predictive of performance on the same authors writing in a different context.

To summarise, in the present study we introduce a feature-set developed by examining measures for three traits that have relationships with political beliefs; social dominance orientation, need for cognitive closure and need for cognition (Cacioppo and Petty, 1982; Pratto et al., 1994; Webster and Kruglanski, 1994). We compare a support vector machines model trained using these features against a vectorised text-only approach and an approach that uses stylistic features chosen without reference to psychological, political or sociological literature. We then test each model on a non-political data set, to try and capture which model is truly classifying based on political stance as opposed to context specific clues. Our aim is to highlight the possibilities that even a light-touch approach to domain-specific knowledge, in this case from the field of psychology, can offer researchers, whilst also offering a potential avenue of research that address model generalisability issues.

2. Related Work

In this section we begin by reviewing work on political affiliation classification, before discussing the importance of considering model generalisability as part of the model testing process. We then introduce our psychologically informed approach to feature selection, citing relevant empirical evidence of traits associated with political affiliation. Finally, we define our experimental aims and hypotheses.

2.1 Political Affiliation Classification

Political affiliation classification can be defined as the task of determining an author's political stance from their written (or oral) communications. Much of the early work in this area focused as classifying authors, often politicians, by membership of a political party (Dahllof, 2012; Diermeier et al., 2012; Yu et al., 2008; Yu and Diermeier, 2010). Historically, researchers used classic machine learning techniques such as Support Vector Machines with 'bag of words' (BoW) feature sets. Feature

selection was performed using formulas such as term-frequency inverse document frequency (TFIDF).

These early efforts yielded promising results. For example, researchers were able to classify congressional speeches correctly as Republican or Democrat in 80% of instances (Yu et al., 2008). Work classifying social media users, particularly Twitter users, also appeared to have good results (Makazhanov and Rafiei, 2013; Pennacchiotti and Popescu, 2011). For example, Joshi and colleagues (2016) gave an accuracy figure of 68% when classifying twitter uses as Republican or Democrat. More recently, researchers have reported accuracies over 90% when classifying tweets (Ullah et al., 2021). Researchers have also been able to classify celebrities by their political affiliation using tweets (Das et al., 2021). Here, features specific to Twitter have been utilised by researchers, including hashtag usage alongside stylistic measures and text vectors. In addition, work has been carried out in various languages, such as Chinese (Gu and Jiang, 2021).

Despite the good results shown in many of these studies, it is rare to see any form of theoretical justification for the features used. Instead, feature sets mostly appear to be decided in a data-driven way, that is based on experimental results or received wisdom in natural language processing practice, rather than based on any empirical evidence of psychological traits known to be associated with political affiliation or relevant theoretical frameworks. In Section 2.3, we discuss the extensive psychological research in this area, which forms the foundation of our methodological approach, detailed in Section 3.

2.2 Generalisability

Whilst the literature provides us with many examples of models exhibiting high accuracy results, improving the generalisability of models across time or topic has not been prioritised. In Psychology, generalisability refers to the ability to extrapolate from findings of a study to the target population at large. However, in this case, generalisability refers to the performance of the model on a different dataset, perhaps collected at a different time, or containing text that covers different topics or is written by a different author, and still achieve good results. A replication of studies that classified Twitter users found that accuracy dropped by as much as 30% when classifiers were used on

everyday users, rather than political figures (Cohen and Ruths, 2013). In this case the model failed to generalise across author-type, as well as topics. This could suggest that the model is using features inherent to political speech to assign class labels, as opposed to some inherent writing style linked to political affiliation. The issue of generalisation has also been addressed in the PAN (a long-running series of tasks and events focusing on text exploration and classification) 2020 task, where fandoms of fan-fiction were varied in an cross-domain authorship verificationtask (Bevendorff et al., 2020).

In a further example that highlights the need to pay attention to generalisability, a model ostensibly trained to classify orators in the Canadian parliament by political party instead appeared to have labelled them by party political status: in or out of power (Hirst et al., 2010). This confound was discovered when the authors applied their model to data collected in a different period of time, to test its generalisability. Perhaps this is why other models of this kind have failed to maintain accuracy across time (Yu et al., 2008). Despite this risk, most models are not tested on datasets that feature data covering different topics or timeframes, meaning it is difficult to discern, from published results, how well these models are likely to generalise.

2.3 Domain-Specific Knowledge

We posit that many of the issues that lead to poor generalisation could be addressed by introducing domain-specific knowledge into the feature selection process. Over-fitting of a model to the training domain occurs because the optimal models tend to be biased towards surface features, such as topical key words, rather than features that are inherent or typical to the domain or attribute of interest. Whilst stylistic and other non-lexical features have been widely used, to our knowledge, previous approaches tend to be data-driven, selecting features based on algorithmic evidence, rather than domain knowledge and theory. It is our objective to explore how introducing foundational domain-specific knowledge can improve model performance. In addition, models often rely on text which contains topics that may only be relevant to a certain type of user or point in time. As an example, post the 2020 United States Presidential election voter fraud became a popular topic amongst Republicans, with as many as 77% of voters for the Republican candidate, Donald Trump, believing this type of fraud was commonplace (Pennycook and Rand, 2021). Words related to voter fraud word be highly useful features therefore for a model training on text written post 2020. However,

the same model might struggle to categorise text written in 2015, when the topic was far less popular. Therefore, there is a need to introduce features free from the influence of topic, that draw upon relevant theory. We suggest a stylistic feature set created with reference to psychological traits correlated with conservatism and liberalism. Stylistic features are commonly defined as features that represent distinctive patterns or trends in an author's writing, rather than the content or topic of the text. Much like authorship identification, stylometry has a long history and has often been used to aid author classification (Holmes, 1998). Examples can include counts or ratios of parts of speech or punctuation usage, with the idea that these features tap into authorship style over content, and are therefore able to tell us something about the 'who' of the author, as opposed to the 'what' of the text content (Kavuri and Kavitha, 2020; Lagutina et al., 2019). Stylistic features focus on pervasive and often unconscious forms of expression and may vary less than content-based features with topic or subject matter. These features should tap into the traits underpinning belief, and therefore allow a model to remain relevant across topic and time. In the case of the present study, we theorised that using stylistic features would improve model generalisability across topic where the authors remained consistent.

Below, a sentence is broken down into parts of speech, a common stylistic feature.

The	quick	brown	fox	jumped	over	the	lazy	dog
determiner	*adjective*	*adjective*	*noun*	*verb*	*preposition*	*determiner*	*adjective*	*noun*

To test this approach we selected three psychological traits of interest due to their relationships to political belief evidenced in the literature. These are Social Dominance Orientation (SDO), Need for Cognitive Closure (NFCC) and Need For Cognition (NFC) (Cacioppo and Petty, 1982; Pratto et al., 1994; Webster and Kruglanski, 1994). To our knowledge, with the exception of one paper limited to the use of nouns and NFCC, there has been no work that has examined how traits linked to political belief might manifest in writing (Cichocka et al., 2016). Therefore, we decided to draw upon traits shown to have relationships with political affiliation, and extrapolate from them. We seek to demonstrate that minimal reference to relevant domain knowledge can improve model performance, even without the costs associated with recruiting participants to create a primary data-set. For this reason, we used measures of these three traits as references to justify feature selection.

Social dominance orientation

Social dominance orientation (SDO) reflects a person's preference for hierarchy. A person scoring high in this trait would prefer for society to be organised in such a way that some groups are higher than others, and they believe that there is a natural order to society (Pratto et al., 1994). SDO has been shown to predict conservatism (Harnish et al., 2018; Pratto et al., 1994; Wilson and Sibley, 2013). Given that conservatism has been defined in the literature as a reluctance to change, a desire to maintain existing order, and an acceptance that society will always be to some extent unequal, the parallels with SDO are clear and it is not surprising that the two are linked (Huntington, 1957; Jost et al., 2003). More recent research suggests that SDO can be seen not only as a preference for hierarchy, but as a strategy for gaining power and maintaining ingroup dominance (Sinn and Hayes, 2018). An example of how this trait might manifest in Republican policy is encapsulated particularly in the anti-immigration policies of the party, such as the so called 'Muslim Ban', where then President Donald Trump prevented residents of several predominantly Muslim countries from entering the United States of America (ACLU, 2017). This policy fits neatly with research which found SDO to be strongly associated with low warmth towards immigrants, as well as anti-immigration attitudes (Satherley and Sibley, 2016). In our approach, we relied on a measure of Social Dominance Orientation developed by Ho and colleagues (2015) for the present study (appendix 1)

Need for cognitive closure

Need for cognitive closure (NFCC) (Webster and Kruglanski, 1994) reflects an individual's preferences and motivations for making judgments and interpreting information. Those high in the trait seek quick answers to questions and dislike ruminating on an issue. They feel uncomfortable when faced with ambiguity, and conversely comfort when given certainty. Once they have found an answer, they are resistant to change even if their view is proven to be factually inaccurate (Kruglanski, 1996). Need for cognitive closure has been shown to be higher in those with conservative views; indeed, a meta-analysis conducted by Jost and colleagues (2003) found that need for cognitive closure correlated significantly with self-reported conservatism. We used another short-from measure (Roets and Van Hiel, 2011) for inspiration, and again relied on sub-facets as well as individual questions (see appendix 2 For scale).

Need for Cognition

Need for cognition (NFC) can be summarised as a drive to think deeply about and fully comprehend a subject or problem (Cacioppo and Petty, 1982). Those high in this trait enjoy exploring the facets of an argument, in almost direct contrast to those high in need for cognitive closure. For example, a person high in NFC would report putting more effort into thinking about a task, and also recall multiple argument messages post-task (Cacioppo et al., 1983). Need for cognition has been found to be positively correlated with liberal views and attitudes, and negatively correlated with conservativism; however, it is important to note that the correlation, whilst significant, is small (Ksiazkiewicz et al., 2016). As with the scales used above, we use a short form version of an original scale, namely the six-item need for cognition scale developed by Lins de Holanda Coelho and colleagues (2018) (see appendix 3 for scale).

In summary, we posit that we can map from the kinds of traits identified above to specific stylistic features and that the features inspired by these traits should be similarly present, and discriminatory, in both political and non-political writing, as the traits themselves remain consistent across time and setting when topics do not. We therefore expected models containing such features would generalise better than models that did not.

Following the above, we developed the following formal hypotheses:

H1 The text-only model will be the weakest performer on the test set.

H2 The model trained using theory informed features will outperform the non-theory informed feature-set.

2.4 Experimental Design

To determine the effectiveness of this approach a modified testing approach is required. Work in the field of data science often relies on results of k-fold cross-validation as a measure of performance or hold-out test set performance, and in particular cross validation is favoured when datasets are relatively small as in the present study (Yadav and Shukla, 2016). Figure 1 describes the process of k-fold cross validation, where results are given as the mean of performance across the various folds.

Figure 2, in contrast, shows a hold-out test set approach, where a model is trained on the training set alone and then performance is measured on the unseen test set. However a hold-out test set is typically drawn from the same domain as the training data and therefore is likely to contain the same topical characteristics. Both approaches, we argue, run the risk of the model being over-fitted to the idiosyncratic properties of the training data, rather than the attribute domain itself, which can result in poor generalisability of the model.

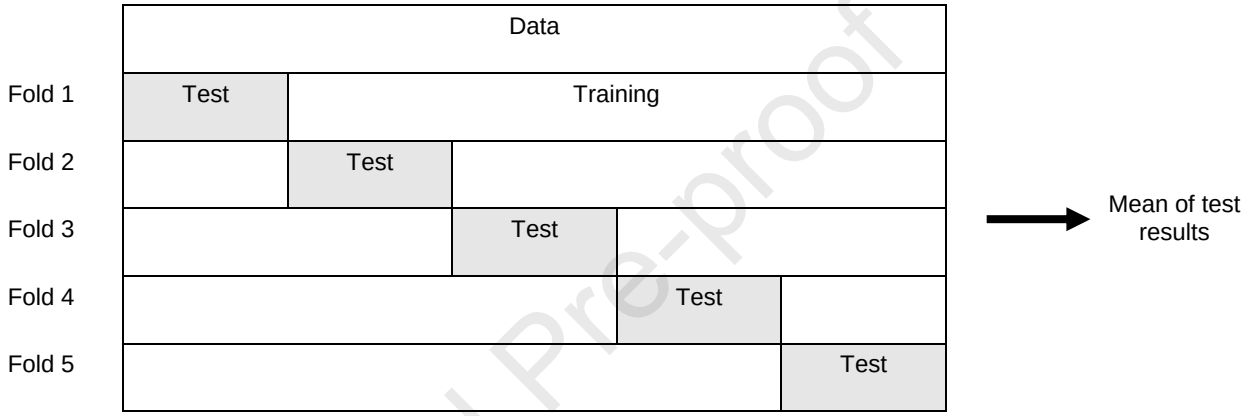

Figure 1. A representation of k-fold cross validation. Results are calculated by finding the mean performance on each fold.

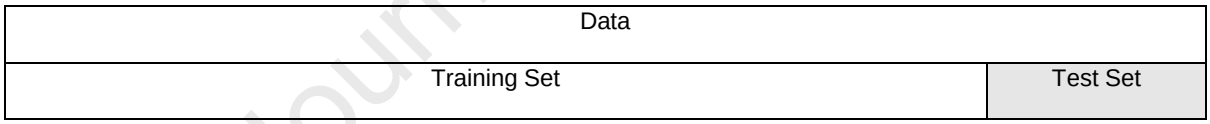

Figure 2. Holdout test set. Here performance on the test set, which is unseen by the model during training, is reported.

To address this problem, we applied a dataset that features non-political speech, posted by the same authors, as a test set. The test set unseen by the model during training and does not contain the same topics or themes as the training data. In this way we hope to provide a better assessment of generalisability, in a similar fashion to the approach taken at PAN 2021(PAN, 2021). We compare models trained using a text-only approach, a text and standard stylistic feature approach, and a model trained using our domain specific stylistic features and text. The aim is to explore how synthesising knowledge from different fields (in this case author profiling and psychology), can be of benefit to data scientists. In all other aspects, we try to use standard practices in the field to investigate the impact of just the addition of the psychologically informed features.

243

3. Methodology

Dataset Creation

Two datasets were created by collecting posts from the Reddit API and a python script using the PRAW wrapper (Boe, 2016). The first dataset was made up of posts made in the r/democrats and r/republican subreddit. All authors with more than four posts in the dataset were retained, however in the case of a user having made fewer posts, if they had more downvoted than upvoted posts they were dropped. This was to filter out potential troll posters who might post infrequently to elicit a negative reaction. Comments were then concatenated by author, and only those who had written more than 100 words were kept. A flow chart of these steps can be seen in figure 3.

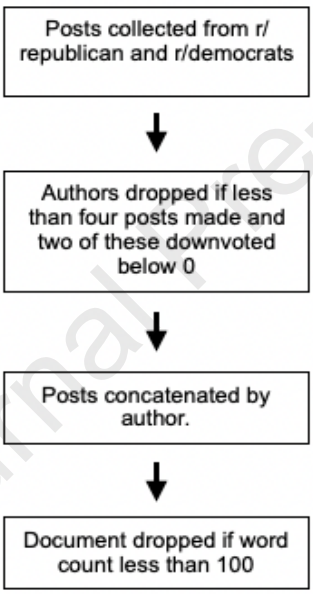

Figure 3. Diagram to show dataset creation

To create a second dataset made up of non-political posts, we collected posts made by the authors in the original dataset in other subreddits: r/amitheasshole, where users ask Redditors to provide feedback on morally ambiguous situations, and r/todayilearned, where users share interesting knowledge they have learned. These subreddits were chosen as there was a large number of users in the political dataset who had made posts in them. Again, documents were concatenated by author and dropped if they were less than 100 words in length. This gave us 811 Democrat authors and 424 Republican authors in this non-political dataset. To balance the classes, we dropped half of the

Democrat authors at random, giving us a new total of 406. The sample method included with the Pandas library was used.

Table 1 below shows key information for the training dataset. The mean comment length was 60 words (0dp) and the mean document (concatenated comments) length was 452 (0dp). Some authors were super-contributors; 14 users made over 200 comments, and two made over 1000. The max number of posts was 2887.

Table 1. Training dataset basics

Subreddit	Number of Authors/Documents	Number of Posts	Number of Documents in Non-Political Dataset	Mean Number of Posts per Author (2dp)	Mode Number of Posts per Author
r/democrats	4,366	50,751	811	11.62	3
r/republican	4,242	44,157	424	10.41	3
Total	8,608	94,908			

Across both datasets, Democrat authors were coded as 0, and Republican authors were coded as 1.

3.1 Feature-set Development

The three measures chosen were the Social Dominance Orientation Short Scale (Ho et al., 2015) (appendix 1), the short form Need for Cognitive Closure Scale (Roets and Van Hiel, 2011) (appendix 2), and the Need for Cognition Scale (Lins de Holanda Coelho et al., 2018) (appendix 3). A detailed list of all features, relevant trait and extraction method can be found in appendix 4.

Table 2 shows all features intended to tap into Need for Cognition. In keeping with the statements in the measure shown in appendix 3, we tried to select features that would convey a sense of openness and complex thought. For example, we scored posts using several measures of readability, as we hypothesised that a higher level of writing might indicate more complex thinking and argumentation.

Table 2. Features inspired by Need for Cognition

Features			
Mean comment length	Dash	Conjunctions	Dale-Chall
Mentions of subreddits	Flesch-Kincaid Grade Level	Question mark	Mean syllables per sentence
Mentions of users	Gunning-Fog	Colons	Number of hapax legomena
Pronouns	Automated Readability Index	Semicolons	Average sentence length
Urls and Emails	Coleman-Liau Index	Commas	Average characters per sentence

Table 3 shows features intended to map to Need for Cognitive Closure. Here we tried to capture the sense of certainty craved by individuals high in this trait. For example, we selected modal verbs of obligation (must, should etc.), as these have a definite feel to them.

Table 3. Features inspired by Need for Cognitive Closure (* indicates hypothesised negative correlation with trait)

Features			
Nouns	First person plural pronouns	Proper Nouns	Adverbs of certainty high*
Possessive Nouns	Third person plural pronouns	Exclamation mark	Adverbs of certainty low
Determiners	Modal verbs of obligation	Modal verbs of possibility*	Adverbs of frequency high*

Table 4 sets out the features linked to Social Dominance Orientation. As an example here, third person plural pronouns such as “they” were intended to map onto the desire for group divisions.

Table 4. Features inspired by Social Dominance Orientation (* indicates hypothesised negative correlation with trait)

Features	
Money	Comparative adjectives
Possessive pronouns - first person singular	Superlative adjectives
Possessive pronouns	Emojis*

- first person plural	
Possessive pronouns	Smileys*
- third person singular	
Possessive pronouns	Possessive pronouns
- third person plural	- second person*

Further examples of features linked to each construct are given in the next section to illustrate extraction techniques.

Feature Extraction and Data Preparation

A variety of techniques were used for feature extraction. Where feasible, Python code was used to calculate word counts. For slightly more complex extraction, such as counts of types of punctuation, regular expressions were used in a script written by the authors. At the level above this, we made use of prebuilt Python libraries. For example, we used the Readability library (Cranenburgh, 2019) for Gunning-Fog scores, Automated Readability Index, and Flesch-Kincaid grade-level measure.

In order to obtain parts of speech counts we used two separate parts of speech taggers: TweetNLP (Owoputi et al., 2013) and the NLTK parts of speech taggers (Bird et al., 2009). TweetNLP deals well with slang and the short posts made on social media, however the NLTK tagger provided extra tags such as determiners. For a full list of all features and the extraction techniques used, see appendix 4. All non-text features were normalised using the SciKit Learn (Pedregosa et al., 2011) libraries normalise function which applies L2 normalisation (values are scaled so that the sum of squares is 1). Normalisation improves performance when using a distance-based method such as SVM (Ali and Smith-Miles, 2006).

Below is a brief description of a selection of features, to illustrate our rationale as well as the extraction process.

Nouns

Work by Cichocka and colleagues (2016) found that conservatives prefer nouns over adjectives. It is hypothesised that individuals scoring higher on NFCC prefer to use nouns over adjectives as a way of defining or stereotyping people or other entities.

"Small fact1, though the Navajo2 wind3 talkers4 was the biggest group5, they weren't the only group6 of natives7 who used their language8 to create an unbreakable code9."

A sentence from the non-political data set with nouns highlighted.

In the above example, nouns are highlighted in yellow. The taggers used take into account the context of the word to estimate the correct part of speech. The final figure for this piece of text would be $\frac{n}{w}$, where n is number of nouns and w is total number of words in the document.

Mentions of Subreddits

As those high in need for cognition prefer more complex thinking, we searched for mentions of subreddits in posts. The idea here is that referencing other sources is a more complex form of argumentation. We used a regular expression to extract these features.

Tables 5 and 6 set out how this process works. Again, the final figure is found by dividing the number of subreddit mentions by the total number of terms in the document.

Table 5. Breakdown of regex phrase

Regular Expression **\sr/.+**	**Matches**
\s	Any whitespace
r/	"r/"
.	Any single character
+	One or more of the preceding item

Table 6. Matches to the Regex phrase

Test Phrase	Match
r/test	Yes
r/1test	Yes
r/!test	Yes
/r/test	No
rtest	No

Superlative adjectives

We hypothesised that an individual high in SDO and, more specifically, the dominance sub-facet, may tend to make more comparisons and seek to define things and other groups as better or worse than, because comparisons allow them to define one groups as dominant and another as subservient. We therefore added comparative adjectives to the feature set. Here again, a part-of-speech tagger was used. Below is an example of a post from the non-political test set with the comparative adjectives highlighted.

"I'm sure what you said is very true, and it's made complicated by the fact that some American products do have better quality. I work for a manufacturer with operations in both the U.S. and overseas. The products sold overseas are sold under a different brand, worse quality, and are cheaper because that's what the people there want. Americans expect higher quality, so that's what they get (along with a higher price). Knowing which products are better (and by how much)... that's a tricky question."

Again, a final figure is found by dividing the number of comparative adjectives in a document by the total number of terms.

3.2 Text Preparation

In order to make the word terms useful as features, they must be represented numerically. We did this using the following standard pre-processing steps:

1. Stopword removal
2. Lemmatization
3. TF-IDF vectorization

Stop word Removal

This is the process of removing words that do not carry meaning and are very common and therefore unlikely to be useful for modelling. We used the list of stop words that comes as a part of the NLTK python library. There are 179 words altogether, and examples include “it”, “am” and “is”.

Lemmatization

This refers to reducing words with the same basic root meaning to one form. An example of this is shown in Table 7.

Word	Reduced Form
Loudly	
Louder	Loud
Loudest	

Table 7. Three words and the lemmatized output

TF-IDF Vectorization

This is a method of numerically representing every word in a corpus (collection of documents). The below formula is used to give each term a score that represents how important it is.

$$TF(t(term\ of\ interest), d(document) = \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d}$$

$$IDF(t) = \log \frac{total\ number\ of\ documents\ in\ corpus}{1 + number\ of\ documents\ containing\ t}$$

$$TF - IDF = TF * IDF$$

3.3 Modelling

We trained our models using a support vector machine with a linear kernel as it is relatively simple to understand and a common approach in the field. Figure 7 shows a basic depiction of an SVM model, where the aim is to find the optimal hyperplane, where the distance between the hyperplane and the closest data points, or support vectors, is maximised.

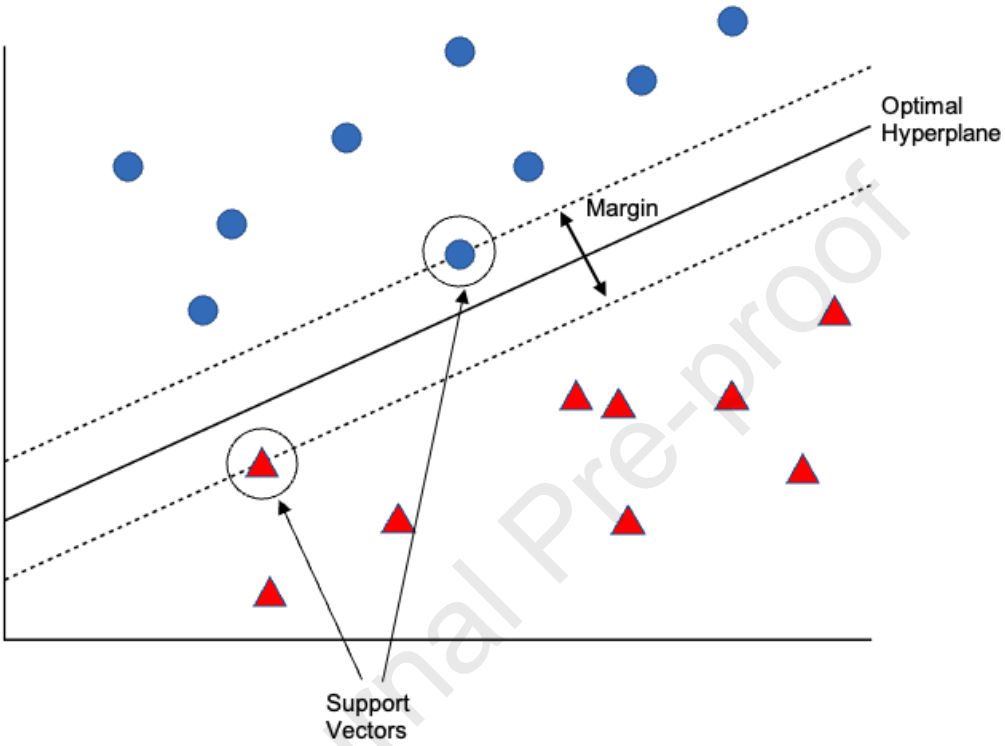

Figure 4. Graphical depiction of the basic principles of the support vector machine algorithm.

As the present study did not specifically seek to maximise performance but instead demonstrate the impact of feature-set, we tuned for C and performed no feature selection. C is an optimization parameter that effects the size of the margin in the model. A larger C will give a smaller margin, and a smaller C a larger margin, as shown in Figure 8. We used the gridsearch feature in SciKit Learn, which inputs multiple values of given parameters and uses cross validation to determine the best performer, to find C for each model type. A C of 1 was selected for the text only and random stylistic feature model, whereas 0.1 was selected for the theory driven dataset.

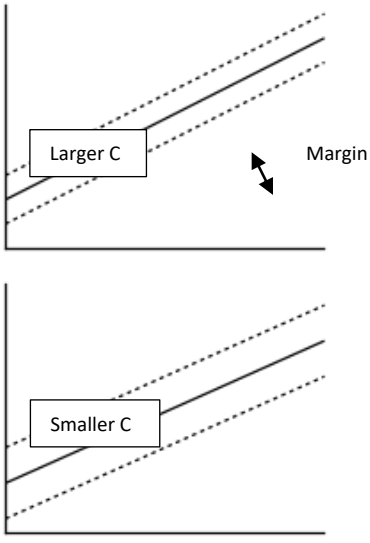

Figure 5. Hyperplanes and margins with differing values of C

The output of the model is a label of Republican or Democrat for each author in the dataset, based on the inputs or feature-set.

4. Results

This study sought to examine the usefulness of stylistic features, informed by psychological theory as a guard against poor generalisability in text classification. Having created a series of models using multiple feature sets, we present our findings below. We use performance on the non-political data set as an indicator of a model's generalisability across topics.

We use accuracy, which is the percentage of authors assigned the correct label, as our main performance metric as the test set was balanced. However, we also report F1, as this is commonly used in classification tasks. This is the harmonic mean of Recall and Precision and is preferred when a dataset is unbalanced. The lower the score, the poorer the performance. An F1 of 1 would be considered perfect performance. F1 is calculated for each class. Here we report the mean F1 score for both classes.

$$F1 = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

$$Precision = \frac{true\ positives}{true\ positives + false\ positives}$$

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

Table 8 shows the results during training on the political posts (5-fold cross validation average) as well as performance on the non-political test set.

Features Used	Accuracy (%)		F1	
	Cross Validation	Non-Political Test Set	Cross Validation	Non-Political Test Set
Theory Driven Features and Text	80.89%	53.86%	0.801	0.538
Text Only	81%	52.77%	0.81	0.528
Random Stylistic and Text	81.60%	51.08%	0.816	0.51

Table 8 -Table of results for models on cross validation and non-political test

During training, the model that used random stylistic features and text is the better performer with an accuracy of 81.6%. This is a similar result to previous work in the area and is not surprising. This feature set contained additional stylistic features that may map onto political affiliation or speech in a way we did not explore.

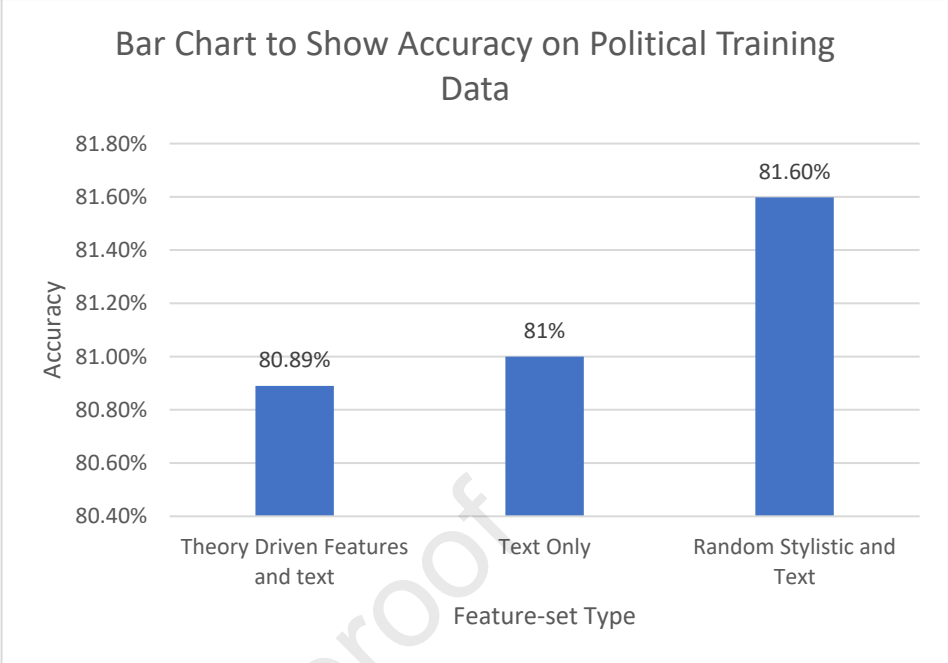

Figure 6 – A bar chart of accuracy scores (%) on training data

However, as shown in figure 10, the performance of all the models drops when tested on the non-political dataset. The model that includes our psychologically informed features suffers from the smallest drop in performance, outperforming both of the other models, albeit by a small margin. The model that was previously the best performer is now the worst.

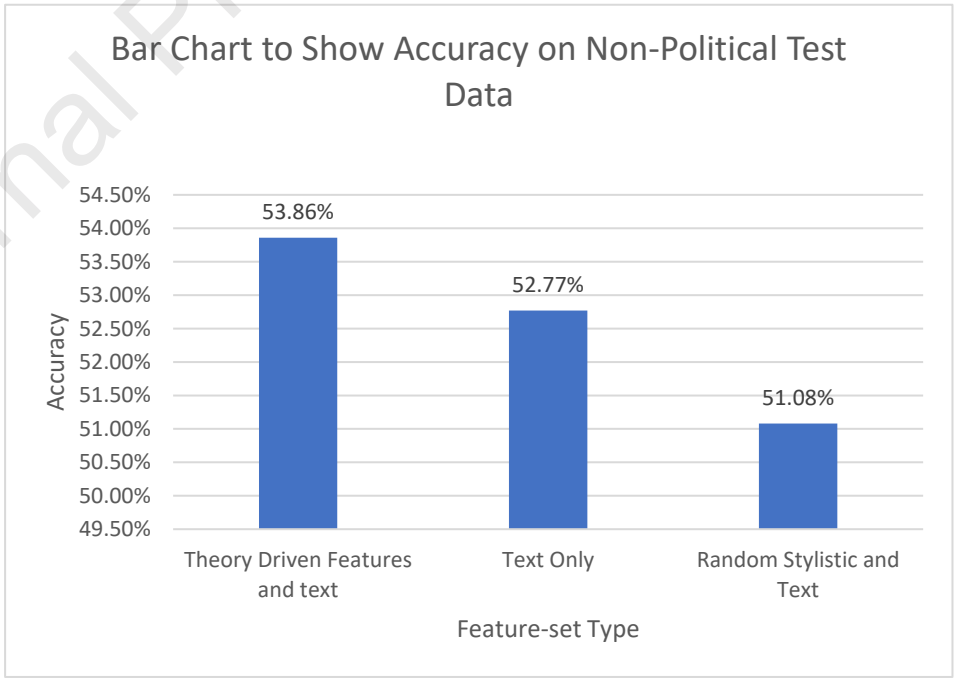

Figure 7 – A bar chart of accuracy scores (%) on test data

In addition to the modelling, we carried out t-tests to examine differences between the two groups for the stylistic markers of each trait. Variables were reverse scored where they were predicted to be negatively associated with conservatism (see appendix 1 for details). We then z-scored the variables and summed all the variables associated with each trait for every author in the dataset. This gave us

an overall score for NFCC markers, NFC markers and SDO markers. Republican authors (*M = -0.026, SD = 3.591*) and Democrat authors (*M = 0.025, SD = 3.469*) were not significantly different on markers of NFCC, $t(8,606) = -0.678$, $p = 0.498$. Similarly, Republicans (*M = 0.02, SD = 7.808)* and Democrats (*M = -0.02, SD = 8.099)* did not differ significantly on markers of NFC, $t(8,606) = -0.234$, $p = 0.815$. However, Republican authors (*M = -0.122, SD = 3.277)* and Democrat authors (*M = 0.119, SD = 3.015)* did differ significantly on markers of SDO, $t(8,500.152) = -3.551$, $p = 0.000$. In the case of the SDO variables, Levene's test was significant ($p = 0.03$) which is why Welch's t-test was used.

5. Discussion

The results of the study show that a feature set created with domain-specific knowledge, in this case psychological traits linked to political affiliation, resulted in small but measurable gains in model generalisability. The model trained using the psychologically informed stylistic set outperformed both the text only model as well as the model that had the added benefit of non-informed stylistic features. We argue that this suggests that the performance gain is not merely an artifact of the use of stylistic features but is in fact linked to the knowledge behind the features. By using a minimal approach that did not involve collecting primary data, we show that while this type of data may be preferable in many ways, it is not necessary for performance gains, lowering the bar in terms of accessibility to researchers from the field of data science. However, future work should seek to optimise the selection of such features through a combination of theory and experimental feedback.

In addition, the performance fall observed on the non-political test set calls into question claims made by authors such as Diermeier and colleagues (2012) that models were sorting authors by underlying ideology; if that were the case, the model should continue to detect ideology in the non-political text. Our findings support the idea that models may be categorising authors based on unknown confounding variables or may be overfitted to overtly political speech (Cohen and Ruths, 2013; Hirst et al., 2010). Furthermore, the results raise questions about the usefulness of stylistic features chosen without reference to theory as a means to improve generalisability. Whilst it is true that these types of features do improve results, it is possible they could be further enhanced, and with relatively low cost to researchers.

Some may argue that models should be trained to achieve the highest accuracy possible, with generalisability less of a concern. We would argue that both objectives are equally important, and that we must think carefully about how models are to be used. If the goal is to create a model that performs as well as possible on one dataset, then the traditional approach is appropriate. On the other hand, if we want to create a model that will generalise across time and topic, we believe it would be sensible for researchers to introduce domain specific knowledge and to also use an alternative test-set, as has been done in other fields (Yin and Zubiaga, 2021). Whilst machine learning can deliver impressive results, there is value in understanding relevant theory, as shown in our results. In

addition, whilst our feature-set was not the best performer on the training data, it cannot be said to be a poor performer. With greater tuning or the use of other modelling techniques any penalisation could be minimised further.

We found no significant differences between r/Republican authors and r/Democrat on the features we tied to NFC and NFCC, whilst the difference between scores for the SDO features was significant. This finding is in contrast to the several studies in psychological literature, including recent work that found that cognitive style was a better predictor of ideological preference than demographic predictors, (Chirumbolo et al., 2004; Jost et al., 2003; Ksiazkiewicz et al., 2016; Zavala et al., 2010; Zmigrod et al., 2021). However, this result is congruent with work in the field of political science suggesting that conservatives and liberals are fairly close cognitively. For example, there is little to separate conservatives and liberals when it comes to physiological response to threats, disgust sensitivity, susceptibility to fake news or, perhaps most intriguingly, the cognitive precedents of populist attitudes (Bakker et al., 2020; Clifford et al., 2022; Erisen et al., 2021; Strandberg et al., 2020). In addition, those higher in political knowledge have been found to be more likely to engage in cognitively complex thinking when evaluating statements incongruent with their political beliefs (Erisen et al., 2018). Given that the members of a political subreddit would almost certainly have more political knowledge than most, this perhaps also explains why the r/Republican members would have little to distinguish them from the r/Democrat users in terms of our two measures of cognitive complexity.

Again, we would suggest that a more in-depth exploration of the stylistic features linked to these traits would be useful here, to rule out the possibility that these findings were merely the result of poorly chosen features. For example, examples of text with corresponding scores on the relevant measures. Although are approach is lightweight and low cost, without this kind of analysis it is too much like guesswork. Having features selected in this way may also improve the model performance overall, as our gains were very minimal.

Furthermore, the mean score on our measure of SDO were higher for the r/Democrat authors, which is in direct contrast to well-established precedents in the literature (Harnish et al., 2018; Wilson and

Sibley, 2013). Here there are two possibilities. It could be that the features selected were not tapping SDO but rather some other unknown variable. Or perhaps party opposition status played a role here as it has done in past analyses (Hirst et al., 2010). Data was collected in mid to late 2020, which meant that redditors in r/Republican had their chosen leader in the White House and also controlled the Senate, as well as having a majority of conservative judges in the Supreme court. In contrast, Democrats only controlled the House, and detested Donald Trump (in fact, the top reason Biden supporters gave for voting for him was that he was not Trump (Atske, 2020). We posit that this context may have meant r/Democrat members felt a sense of continuous threat to their ingroup as they discussed and evaluated Republican policies, which led to traits of SDO being expressed in their writing. They had a powerful outgroup to rail against, whilst the r/Republicans users were in a position of power. Indeed, intergroup threat perception and racism were both found to be higher following manipulation of threat perception, with SDO acting as a moderator (Uenal et al., 2021). Perhaps here a similar effect is occurring, with the threat of Republican dominance increasing the expression of certain stylistic features moderated by SDO. This theory could be tested by collecting posts made in both subreddits since the election of President Biden, and observing the differences in SDO for any change.

5.1 Future Directions

In terms of future work, the introduction of more complex modelling techniques would be a logical extension of this work. In this study, we took a very simplistic high-level approach as we were concerned with showing the usefulness of our approach, rather than developing a state-of-the-art model. We chose an SVM model as that was the approach used by early researchers in the field (Diermeier et al., 2012; Yu et al., 2008; Yu and Diermeier, 2010). We also did not tune any of the model parameters apart from C, again to keep the methodology as simple as possible. However, random forests, logistic regression, naïve bayes, and KNN are all commonly used algorithms for text classification (Pranckevičius and Marcinkevičius, 2017; Shah et al., 2020). Therefore, it would be prudent to explore how a feature set such as the one developed here would affect performance in these cases.

In addition, we did not explore the impact of feature selection to our results. We would suggest a feature ablation study, perhaps using SVM recursive feature elimination (RFE) (Sanz et al., 2018). Here, whilst the number of input variables remains greater than two, a model is trained and features are ranked by the weight of their coefficients squared. The feature with the lowest value is dropped and the model is retrained. Once the process is complete, a ranked list of variables is created. Not only would this aid performance as it would allow unhelpful variables to be removed, it could also reveal interesting results relevant to psychologists. For example, if exclamation marks were found to be a highly useful variable, this would raise interesting questions as to why, opening avenues for future experimental work.

To strengthen the interdisciplinary nature of our approach, we would also suggest using primary data to improve the feature development process. This could involve recruiting participants to complete writing tasks and measures of traits of interest. The text they produce could then be explored for any differences that are linked to trait score. Indeed, previous work has been carried out exploring differences in writing style for those high and low in personality traits such as the Big Five and narcissism (Chung and Pennebaker, 2013; Cutler et al., 2021; Stillwell and Kosinski, 2012). This approach could be extended into other domains and traits, with results collated and shared across disciplines for use by researchers of different backgrounds.

Finally, this approach could be explored across languages to demonstrate that its usefulness is not limited to an American context. There is already work that explores classifying authors by political affiliation in multipole languages and we would hope that here too reference to domain specific knowledge would be of use (Abd et al., 2020; Kapočiūtė-Dzikienė et al., 2014; Lapponi et al., 2018).

5.2 Practical Applications

There are also potential practical applications of the present study. In security research, there may be a desire to flag forum users as extremists so they can be tracked online (Ellen and Parameswaran, 2011). Here we can imagine that it would be vitally important that a model tap into an underlying trait and be generalisable across context. In this way, a user could be identified as dangerous regardless of the topics of their posts. This is especially important given that social media plays a role in the

recruitment process for almost 90% of extremists in 2016 (*START*, 2021). Tools that can provide an early warning of such activity to the appropriate security services are of great value (Gaikwad et al., 2021).

However, the practical applications of our methodology also raise important ethical concerns. In this case, by posting in the r/Democrat and r/Republicans subreddits, users are outing their own political affiliation. However, when we work to develop a model that can classify users who post in non-political spaces, are we violating their privacy? The sanctity of the voting booth is enshrined in the universal declaration of human rights (UN, 1948), and if a political party, government, or organization were able to determine a person's political affiliation without their permission, there could be dangerous ramifications. For example, imagine an autocratic regime that imprisons supporters of rival political group: how could an individual stay safe when the regime could determine their political position, just from posts made in non-political spaces? Further to this, is it appropriate to label a person as extremist, with all the associated connotations, if they have not broken the law? Widescale implementation of this kind of methodology could have a chilling effect on free speech. However, given how underprepared the U.K. government, for example, is in terms of tackling issues such as far-right extremism, perhaps there is an argument to made here about the greater benefit for society at large unprepared (Ozduzen et al., 2021).

5.3 Limitations

In terms of the limitations of our methodology, as previously discussed, we used a very simplistic approach that does not make use of the plethora of state-of-the-art techniques available. Again, this was a deliberate choice made to allow the impact of the feature-set to be more clearly understood. It should also be noted that we looked for correlates and predictors of conservatism and liberalism, whilst our dataset is labelled as Republican or Democrat, respectively. We feel confident that these party affiliations are appropriate proxies for the relevant ideologies given that the definitions given by the literature and the policies of the parties are well-matched (Caplan, 2016; Graham et al., 2009; Jost et al., 2003; Saad et al., 2019). However, there are Conservatives and Liberals who do not identify as Republican or Democrat and vice versa. Indeed, a recent Gallup poll (2022) found that 12% of Democrats identified as Conservative, and 4% of Republicans identified as Liberal. The solution here

would be to create a dataset of posts for authors alongside measures of their political ideology, however there would be heavy financial costs associated with this approach.

5.4 Conclusion

Author profiling remains a popular and enduring task for data scientists. The field of political affiliation classification in particular has a long history, stemming from the classification of politicians to more recent work looking at users of social media (Gu and Jiang, 2021; Yu et al., 2008). In the present study, we have attempted to show how considering field-specific knowledge, in this case psychological theory relating to personality traits, can be helpful to political affiliation inference research. This approach could be especially helpful with reference to the increasingly relevant issues of model generalisability, as highlighted by the recent PAN authorship attribution tasks (PAN, 2021). In addition, our results suggest that past work may have been tapping into confounding variables, as previously suggested by other authors (Hirst et al., 2010). The psychologically informed feature-set we developed showed superior performance to the two approaches that did not involve domain-specific knowledge on the task of determining author political affiliation using non-political text. Future work should seek to extend this approach into other topics and using more sophisticated and nuanced methods.

Acknowledgments

We thank the reviewers for their helpful suggestions: their insight was invaluable. We would also like to thank the Editors.

References

Abd DH, Sadiq AT and Abbas AR (2020) Classifying Political Arabic Articles Using Support Vector Machine with Different Feature Extraction. In: *Applied Computing to Support Industry: Innovation and Technology* (eds MI Khalaf, D Al-Jumeily, and A Lisitsa), Cham, 2020, pp. 79–94. Communications in Computer and Information Science. Springer International Publishing. DOI: 10.1007/978-3-030-38752-5_7.

ACLU (2017) Timeline of the Muslim Ban. Available at: https://www.aclu-wa.org/pages/timeline-muslim-ban (accessed 11 June 2021).

Ali S and Smith-Miles KA (2006) Improved support vector machine generalization using normalized input space. In: *Proceedings of the 19th Australian joint conference on Artificial Intelligence: advances in Artificial Intelligence*, Berlin, Heidelberg, 4 December 2006, pp. 362–371. AI'06. Springer-Verlag. DOI: 10.1007/11941439_40.

Atske S (2020) Perceptions of Trump and Biden. In: *Pew Research Center - U.S. Politics & Policy*. Available at: https://www.pewresearch.org/politics/2020/08/13/perceptions-of-trump-and-biden/ (accessed 2 December 2022).

Bakker BN, Schumacher G, Gothreau C, et al. (2020) Conservatives and liberals have similar physiological responses to threats. *Nature Human Behaviour* 4(6). 6. Nature Publishing Group: 613–621. DOI: 10.1038/s41562-020-0823-z.

Bevendorff J, Ghanem B, Giachanou A, et al. (2020) Shared Tasks on Authorship Analysis at PAN 2020. In: *Advances in Information Retrieval* (eds JM Jose, E Yilmaz, J Magalhães, et al.), Cham, 2020, pp. 508–516. Lecture Notes in Computer Science. Springer International Publishing. DOI: 10.1007/978-3-030-45442-5_66.

Bird S, Loper E and Klein E (2009) *Natural Language Processing with Python*. Newton, Massachusetts, USA: O'Reilly Media Inc.

Cacioppo J, Petty R and Morris K (1983) Effects of need for cognition on message evaluation, recall, and persuasion. DOI: 10.1037/0022-3514.45.4.805.

Cacioppo JT and Petty RE (1982) The need for cognition. *Journal of Personality and Social Psychology* 42(1). US: American Psychological Association: 116–131. DOI: 10.1037/0022-3514.42.1.116.

Caplan D (2016) Log Cabin Republicans: GOP Platform the 'Most Anti-LGBT' in Party's History. Available at: https://abcnews.go.com/Politics/log-cabin-republicans-gop-party-platform-anti-lgbt/story?id=40564850 (accessed 4 June 2020).

Chirumbolo A, Areni A and Sensales G (2004) Need for cognitive closure and politics: Voting, political attitudes and attributional style. *International Journal of Psychology* 39(4). Routledge: 245–253. DOI: 10.1080/00207590444000005.

Chung CK and Pennebaker JW (2013) Linguistic Inquiry and Word Count (LIWC). In: *Applied Natural Language Processing*, pp. 206–229. DOI: 10.4018/978-1-60960-741-8.ch012.

Cichocka A, Bilewicz M, Jost JT, et al. (2016) On the grammar of politics—or why conservatives prefer nouns. *Political Psychology* 37(6). Wiley Online Library: 799–815.

Clifford S, Erisen C, Wendell D, et al. (2022) Disgust sensitivity and support for immigration across five nations. *Politics and the Life Sciences*. Cambridge University Press: 1–16. DOI: 10.1017/pls.2022.6.

Cohen R and Ruths D (2013) Classifying Political Orientation on Twitter: It's Not Easy! *Proceedings of the International AAAI Conference on Web and Social Media* 7(1). 1: 91–99.

Cranenburgh A van (2019) readability: Measure the readability of a given text using surface characteristics. Cython, Python. Available at: https://github.com/andreasvc/readability/ (accessed 12 July 2021).

Cutler AD, Carden SW, Dorough HL, et al. (2021) Inferring Grandiose Narcissism From Text: LIWC Versus Machine Learning. *Journal of Language and Social Psychology* 40(2). SAGE Publications Inc: 260–276. DOI: 10.1177/0261927X20936309.

Dahllof M (2012) Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches—A comparative study of classifiability. *Literary and Linguistic Computing* 27(2): 139–153.

Das KG, Patra BG and Naskar SK (2021) Profiling Celebrity Profession from Twitter Data. In: *2021 International Conference on Asian Language Processing (IALP)*, December 2021, pp. 207–212. DOI: 10.1109/IALP54817.2021.9675260.

Diermeier D, Godbout J-F, Yu B, et al. (2012) Language and Ideology in Congress. *British Journal of Political Science* 42(1). Cambridge University Press: 31–55.

Ellen J and Parameswaran S (2011) Machine Learning for Author Affiliation within Web Forums – Using Statistical Techniques on NLP Features for Online Group Identification. In: *2011 10th International Conference on Machine Learning and Applications and Workshops*, December 2011, pp. 100–105. DOI: 10.1109/ICMLA.2011.90.

Erisen C, Redlawsk DP and Erisen E (2018) Complex Thinking as a Result of Incongruent Information Exposure. *American Politics Research* 46(2). SAGE Publications Inc: 217–245. DOI: 10.1177/1532673X17725864.

Erisen C, Guidi M, Martini S, et al. (2021) Psychological Correlates of Populist Attitudes. *Political Psychology* 42(S1): 149–171. DOI: 10.1111/pops.12768.

Gaikwad M, Ahirrao S, Phansalkar S, et al. (2021) Online Extremism Detection: A Systematic Literature Review With Emphasis on Datasets, Classification Techniques, Validation Methods, and Tools. *IEEE Access* 9: 48364–48404. DOI: 10.1109/ACCESS.2021.3068313.

739 Graham J, Haidt J and Nosek BA (2009) Liberals and conservatives rely on different sets of
740 moral foundations. *Journal of Personality and Social Psychology* 96(5): 1029–1046.
741 DOI: 10.1037/a0015141.

742 Gu F and Jiang D (2021) Prediction of Political Leanings of Chinese Speaking Twitter Users.
743 arXiv:2110.05723. arXiv. Available at: http://arxiv.org/abs/2110.05723 (accessed 24
744 September 2022).

745 Harnish R, Bridges K and Gump J (2018) Predicting Economic, Social, and Foreign Policy
746 Conservatism: the Role of Right-Wing Authoritarianism, Social Dominance
747 Orientation, Moral Foundations Orientation, and Religious Fundamentalism. *Current*
748 *Psychology* 37. DOI: 10.1007/s12144-016-9552-x.

749 Hinds J and Joinson AN (2018) What demographic attributes do our digital footprints reveal?
750 A systematic review. *PLOS ONE* 13(11). Public Library of Science: 1–40. DOI:
751 10.1371/journal.pone.0207112.

752 Hirst G, Riabinin Y and Graham J (2010) Party status as a confound in the automatic
753 classification of political speech by ideology. In: *Proceedings of the 10th International*
754 *Conference on Statistical Analysis of Textual Data (JADT 2010)*, 2010, pp. 731–742.

755 Ho AK, Sidanius J, Kteily N, et al. (2015) The nature of social dominance orientation:
756 Theorizing and measuring preferences for intergroup inequality using the new SDO_7
757 scale. *Journal of Personality and Social Psychology* 109(6). American Psychological
758 Association: 1003.

759 Holmes DI (1998) The Evolution of Stylometry in Humanities Scholarship. *Literary and*
760 *Linguistic Computing* 13(3): 111–117. DOI: 10.1093/llc/13.3.111.

761 Huntington SP (1957) Conservatism as an Ideology. *The American Political Science Review*
762 51(2). [American Political Science Association, Cambridge University Press]: 454–473.
763 DOI: 10.2307/1952202.

764 Joshi A, Bhattacharyya P and Carman M (2016) Political Issue Extraction Model: A Novel
765 Hierarchical Topic Model That Uses Tweets By Political And Non-Political Authors. In:
766 *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity,*
767 *Sentiment and Social Media Analysis*, San Diego, California, June 2016, pp. 82–90.
768 Association for Computational Linguistics. DOI: 10.18653/v1/W16-0415.

769 Jost JT, Glaser J, Kruglanski AW, et al. (2003) Political conservatism as motivated social
770 cognition. *Psychological bulletin* 129(3). American Psychological Association: 339.

771 Kapočiūtė-Dzikienė J, Utka A and Šarkutė L (2014) Feature Exploration for Authorship
772 Attribution of Lithuanian Parliamentary Speeches. In: Sojka P, Horák A, Kopeček I, et
773 al. (eds) *Text, Speech and Dialogue*. Lecture Notes in Computer Science. Cham:
774 Springer International Publishing, pp. 93–100. DOI: 10.1007/978-3-319-10816-2_12.

775 Kavuri K and Kavitha M (2020) A Stylistic Features Based Approach for Author Profiling. In:
776 *Recent Trends in Communication and Intelligent Systems* (eds H Sharma, AKS Pundir,

N Yadav, et al.), Singapore, 2020, pp. 185–193. Algorithms for Intelligent Systems. Springer. DOI: 10.1007/978-981-15-0426-6_20.

Kruglanski AW (1996) Motivated social cognition: Principles of the interface. In: *Social Psychology: Handbook of Basic Principles*. New York, NY, US: The Guilford Press, pp. 493–520.

Ksiazkiewicz A, Ludeke S and Krueger R (2016) The Role of Cognitive Style in the Link Between Genes and Political Ideology. *Political Psychology* 37(6): 761–776. DOI: 10.1111/pops.12318.

Lagutina K, Lagutina N, Boychuk E, et al. (2019) A Survey on Stylometric Text Features. In: *2019 25th Conference of Open Innovations Association (FRUCT)*, November 2019, pp. 184–195. DOI: 10.23919/FRUCT48121.2019.8981504.

Lapponi E, Søyland MG, Velldal E, et al. (2018) The Talk of Norway: a richly annotated corpus of the Norwegian parliament, 1998–2016. *Language Resources and Evaluation* 52(3): 873–893. DOI: 10.1007/s10579-018-9411-5.

Lins de Holanda Coelho G, H. P. Hanel P and J. Wolf L (2018) The Very Efficient Assessment of Need for Cognition: Developing a Six-Item Version. *Assessment*. SAGE Publications Inc: 1073191118793208. DOI: 10.1177/1073191118793208.

Makazhanov A and Rafiei D (2013) Predicting political preference of Twitter users. In: *Social Network Analysis and Mining*, Niagara Falls, 2013, p. 193. IEEE. DOI: 10.1007/s13278-014-0193-5.

Oberlander J and Gill AJ (2004) Individual differences and implicit language: personality, parts-of-speech and pervasiveness. *Proceedings of the Annual Meeting of the Cognitive Science Society* 26(26). Available at: https://escholarship.org/uc/item/94c490mq (accessed 16 February 2020).

Owoputi O, O'Connor B, Dyer C, et al. (2013) Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. In: *Proceedings of NAACL 2013*, Atlanta, GA, USA, 2013, p. 11.

PAN (2021) PAN Shared Tasks. Available at: https://pan.webis.de/shared-tasks.html (accessed 7 May 2020).

Pedregosa F, Varoquaux G, Gramfort A, et al. (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12(85): 2825–2830.

Pennacchiotti M and Popescu A-M (2011) Democrats, republicans and starbucks afficionados: User classification in twitter. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 21 August 2011, pp. 430–438. DOI: 10.1145/2020408.2020477.

Pennington J, Socher R and Manning C (2014) GloVe: Global Vectors for Word Representation. Available at: https://nlp.stanford.edu/projects/glove/ (accessed 13 October 2022).

Pennycook G and Rand DG (2021) Research note: Examining false beliefs about voter fraud in the wake of the 2020 Presidential Election. *Harvard Kennedy School Misinformation Review*. DOI: 10.37016/mr-2020-51.

Petersen W (n.d.) Enum Cohrs1 University of Eastern Finland.: 13.

Pranckevičius T and Marcinkevičius V (2017) Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Baltic Journal of Modern Computing* 5(2). DOI: 10.22364/bjmc.2017.5.2.05.

Pratto F, Sidanius J, Stallworth LM, et al. (1994) Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology* 67(4): 741–763. DOI: 10.1037/0022-3514.67.4.741.

Roets A and Van Hiel A (2011) Item selection and validation of a brief, 15-item version of the Need for Closure Scale. *Personality and Individual Differences* 50(1). Elsevier: 90–94.

Saad L (2022) U.S. Political Ideology Steady; Conservatives, Moderates Tie. Available at: https://news.gallup.com/poll/388988/political-ideology-steady-conservatives-moderates-tie.aspx (accessed 25 September 2022).

Saad L, Jones J and Brenan M (2019) Understanding Shifts in Democratic Party Ideology. Available at: https://news.gallup.com/poll/246806/understanding-shifts-democratic-party-ideology.aspx (accessed 4 June 2020).

Sanz H, Valim C, Vegas E, et al. (2018) SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinformatics* 19(1): 432. DOI: 10.1186/s12859-018-2451-4.

Satherley N and Sibley CG (2016) A Dual Process Model of attitudes toward immigration: Predicting intergroup and international relations with China. *International Journal of Intercultural Relations* 53: 72–82. DOI: 10.1016/j.ijintrel.2016.05.008.

Shah K, Patel H, Sanghvi D, et al. (2020) A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. *Augmented Human Research* 5(1): 12. DOI: 10.1007/s41133-020-00032-0.

Sinn JS and Hayes MW (2018) Is Political Conservatism Adaptive? Reinterpreting Right-Wing Authoritarianism and Social Dominance Orientation as Evolved, Sociofunctional Strategies. *Political Psychology* 39(5): 1123–1139. DOI: 10.1111/pops.12475.

START (2021) Profiles of Individual Radicalization in the United States (PIRUS). Available at: https://www.start.umd.edu/data-tools/profiles-individual-radicalization-united-states-pirus (accessed 12 July 2021).

Stillwell DJ and Kosinski M (2012) myPersonality project: Example of successful utilization of online social networks for large-scale social research. *American Psychologist* 59(2): 93–104.

Strandberg T, Olson JA, Hall L, et al. (2020) Depolarizing American voters: Democrats and Republicans are equally susceptible to false attitude feedback. *PLOS ONE* 15(2). Public Library of Science: e0226799. DOI: 10.1371/journal.pone.0226799.

Uenal F, Sidanius J, Roozenbeek J, et al. (2021) Climate change threats increase modern racism as a function of social dominance orientation and ingroup identification. *Journal of Experimental Social Psychology* 97. DOI: 10.1016/j.jesp.2021.104228.

Ullah H, Ahmad B, Sana I, et al. (2021) Comparative study for machine learning classifier recommendation to predict political affiliation based on online reviews. *CAAI Transactions on Intelligence Technology* 6(3): 251–264. DOI: 10.1049/cit2.12046.

Webster DM and Kruglanski AW (1994) Individual differences in need for cognitive closure. *Journal of personality and social psychology* 67(6). American Psychological Association: 1049.

Wilson MS and Sibley CG (2013) Social Dominance Orientation and Right-Wing Authoritarianism: Additive and Interactive Effects on Political Conservatism. *Political Psychology* 34(2): 277–284. DOI: 10.1111/j.1467-9221.2012.00929.x.

Yadav S and Shukla S (2016) Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. In: *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, February 2016, pp. 78–83. DOI: 10.1109/IACC.2016.25.

Yin W and Zubiaga A (2021) Towards generalisable hate speech detection: a review on obstacles and solutions. *arXiv:2102.08886 [cs]*. Available at: http://arxiv.org/abs/2102.08886 (accessed 25 March 2021).

Yu B and Diermeier D (2010) A longitudinal study of language and ideology in congress. In: *The 68th National Conference of Midwest Political Science Association*, Chicago, IL, 2010.

Yu B, Kaufmann S and Diermeier D (2008) Classifying Party Affiliation from Political Speech. *Journal of Information Technology & Politics* 5(1). Routledge: 33–48. DOI: 10.1080/19331680802149608.

Zavala AGD, Cislak A and Wesolowska E (2010) Political Conservatism, Need for Cognitive Closure, and Intergroup Hostility. *Political Psychology* 31(4): 521–541. DOI: 10.1111/j.1467-9221.2010.00767.x.

Zmigrod L, Eisenberg IW, Bissett PG, et al. (2021) The cognitive and perceptual correlates of ideological attitudes: a data-driven approach. *Philosophical Transactions of the Royal Society B: Biological Sciences* 376(1822). Royal Society: 20200424. DOI: 10.1098/rstb.2020.0424.

Appendix

Appendix 1

Statements found on the Social Dominance Orientation short scale (Ho et al., 2015)

1.	An ideal society requires some groups to be on top and others to be on the bottom.
2.	Some groups of people are simply inferior to other groups.
3.	Groups at the bottom are just as deserving as groups at the top. **
4.	No one group should dominate in society. **
5.	Group equality should not be our primary goal.
6.	It is unjust to try to make groups equal.
7.	We should do what we can to equalize conditions for different groups. **
8.	We should work to give all groups an equal chance to succeed. **

*Table 9. – Items on the short Social Dominance Orientation scale(** indicates reverse scoring)*

Key: Dominance – Yellow, Antiegalitarianism - Blue

Appendix 2

Statements found on the short form Need for Cognitive Closure scale (Roets & Van Hiel, 2011)

1.	I don't like situations that are uncertain.
2.	I dislike questions which could be answered in many different ways.
3.	I find that a well-ordered life with regular hours suits my temperament.
4.	I feel uncomfortable when I don't understand the reason why an event occurred in my life.
5.	I feel irritated when one person disagrees with what everyone else in a group believes.
6.	I don't like to go into a situation without knowing what I can expect from it.
7.	When I have made a decision, I feel relieved.
8.	When I am confronted with a problem, I'm dying to reach a solution very quickly.
9.	I would quickly become impatient and irritated if I would not find a solution to a problem immediately.
10.	I don't like to be with people who are capable of unexpected actions.
11.	I dislike it when a person's statement could mean many different things.
12.	I find that establishing a consistent routine enables me to enjoy life more.
13.	I enjoy having a clear and structured mode of life.
14.	I do not usually consult many different opinions before forming my own view.
15.	I dislike unpredictable situations.

Key for Facets

Facet	Colour
Order	
Predictability	
Decisiveness	
Ambiguity	
Closed-mindedness	

Table 10. – Items on the short form Need for Cognitive Closure scale

Appendix 3

Statements found on the six item Need for Cognition Scale (Lins de Holanda Coelho et al., 2018)	
1.	I prefer complex to simple problems.
2.	I like to have the responsibility of handling a situation that requires a lot of thinking.
3.	Thinking is not my idea of fun.**
4.	I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.**
5.	I really enjoy a task that involves coming up with new solutions to problems.
6.	I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.

*Table 11. – Items on the short 6 item need for cognition scale (** indicates reverse scoring)*

Appendix 4

Feature	Related To	Normalised by Word Count	Relationship to Republicanism	Extraction Technique
Mean comment length	Need for Cognition	No	Negative	Author created code
Mentions of subreddits	Need for Cognition	Yes	Negative	Regular Expression
Mentions of users	Need for Cognition	Yes	Negative	Regular Expression
Pronouns	Need for Cognition	Yes	Negative	PoS Tagger[1]
Urls and Emails	Need for Cognition	Yes	Negative	Regular Expression
Conjunctions	Need for Cognition	Yes	Negative	PoS Tagger
Question mark	Need for Cognition	Yes	Negative	Regular Expression
Colons	Need for Cognition	Yes	Negative	Regular Expression
Semicolons	Need for Cognition	Yes	Negative	Regular Expression
Commas	Need for Cognition	Yes	Negative	Regular Expression
Dash	Need for Cognition	Yes	Negative	Regular Expression
Flesch-Kincaid Grade Level	Need for Cognition	No	Negative	Readability library
Gunning-Fog	Need for Cognition	No	Negative	Readability library
Automated Readability Index	Need for Cognition	No	Negative	Readability library
Coleman-Liau Index	Need for Cognition	No	Negative	Readability library
Dale-Chall	Need for Cognition	No	Negative	Readability library
Mean syllables per sentence	Need for Cognition	No	Negative	Syllapy library
Number of hapax legomena	Need for Cognition	Yes	Negative	NLTK library
Average sentence length	Need for Cognition	No	Negative	Author created code
Average characters per sentence	Need for Cognition	No	Negative	Author created code

[1] TweetNLP and NLTK parts of speech taggers used

Nouns	Need for Cognitive Closure	Yes	Positive	PoS tagger
Possessive Nouns	Need for Cognitive Closure	Yes	Positive	PoS tagger
Determiners	Need for Cognitive Closure	Yes	Positive	PoS tagger
Proper Nouns	Need for Cognitive Closure	Yes	Positive	PoS tagger
Exclamation mark	Need for Cognitive Closure	Yes	Positive	Regular Expression
First person plural pronouns	Need for Cognitive Closure	Yes	Positive	Regular Expression
Third person plural pronouns	Need for Cognitive Closure	Yes	Positive	Regular Expression
Modal verbs of obligation	Need for Cognitive Closure	Yes	Positive	Regular Expression
Modal verbs of possibility	Need for Cognitive Closure	Yes	Negative	Regular Expression
Adverbs of certainty high	Need for Cognitive Closure	Yes	Positive	Regular Expression
Adverbs of certainty low	Need for Cognitive Closure	Yes	Negative	Regular Expression
Adverbs of frequency high	Need for Cognitive Closure	Yes	Positive	Regular Expression
Money	SDO	Yes	Positive	Regular Expression
Possessive pronouns - first person singular	SDO	Yes	Positive	Regular Expression
Possessive pronouns - first person plural	SDO	Yes	Positive	Regular Expression
Possessive pronouns - third person singular	SDO	Yes	Positive	Regular Expression
Possessive pronouns	SDO	Yes	Positive	Regular Expression

- third person plural				
Comparative adjectives	SDO	Yes	Positive	PoS tagger
Superlative adjectives	SDO	Dominance	Positive	PoS tagger
Emojis	SDO	Dominance	Negative	Emojis library
Smileys	SDO	Dominance	Negative	Regular Expression
Posessive pronouns - second person	SDO	Yes	Negative	Regular Expression

Table 12 – Feature-set breakdown

Supplementary Materials

Feature	Number of non-zero points in corpus	Mean	Median	Mode
Mentions of subreddits	503	0.1	0	0
Mentions of users	53	0.01	0	0
Pronouns	8606	37.63	20	10
Urls and Emails	1947	0.77	0	0
Conjunctions	8608	77.56	39	21
Question mark	6041	3.39	2	0
Colons	2404	0.72	0	0
Semicolons	982	0.26	0	0
Commas	8206	15.08	7	0
Dash	4185	2.32	0	0
Nouns	8608	85.48	44	24
Possessive Nouns	1202	0.21	0	0
Determiners	8607	47.25	24	12
Proper Nouns	8198	16.2	8	0
Exclamation mark	2882	1.33	0	0
First person plural pronouns	5888	3.31	1	0
Third person plural pronouns	7108	5.63	3	0
Modal verbs of obligation	6408	3.19	2	0
Modal verbs of possibility	4208	1.24	0	0
Adverbs of certainty high	5104	1.67	1	0
Adverbs of certainty low	1800	0.35	0	0
Adverbs of frequency high	3645	0.97	0	0
Money	661	0.19	0	0

Possessive pronouns - first person singular	3933	1.2	0	0
Possessive pronouns - first person plural	2723	0.73	0	0
Possessive pronouns - third person singular	4914	2.19	1	0
Possessive pronouns - third person plural	4724	1.66	0.99	0
Comparative adjectives	5170	1.86	1	0
Superlative adjectives	4110	1.14	0	0
Emojis	569	0.23	0	0
Smileys	351	0.06	0	0
Posessive pronouns - second person	4049	1.37	0	0

Table 13 – Feature-set breakdown

Feature	Examples
Mentions of subreddits	e.g. r/amitheasshole, r/todayilearned, r/news
Mentions of users	e.g. u/jonesmith
Pronouns	All pronouns e.g. him, her, me, mine
Urls and Emails	e.g. www.example.com, example@example.com
Conjunctions	e.g. and, for, yet
Question mark	?
Colons	:
Semicolons	;
Commas	,
Dash	-
Nouns	Any noun
Possessive Nouns	Any possessive noun e.g. the researcher's example
Determiners	e.g. which, whether
Proper Nouns	Any proper noun e.g. Queen
Exclamation mark	!
First person plural pronouns	us, we
Third person plural pronouns	they, them
Modal verbs of obligation*	should, must, need to, will, had better, shall, ought, have to
Modal verbs of possibility*	may, maybe, might, could, cud
Adverbs of certainty high*	certainly, clearly, definitely, doubtless, indeed, emphatically, likely, always, unlikely, obviously, invariably, absolutely, presumably, really, surely, truly, evidently, unquestionably
Adverbs of certainty low	possibly, maybe, perhaps
Adverbs of frequency high	always, often, usually, never, generally, rarely, seldom, hardly
Money	$, £, ¥,€
Possessive pronouns - first person singular	mine, my

Possessive pronouns - first person plural	ours
Possessive pronouns - third person singular	his, hers, its
Possessive pronouns - third person plural	theirs
Comparative adjectives	Any comparative adjective e,g, larger, smaller
Superlative adjectives	Any superlative adjective e,g, biggest, smallest
Emojis	Any emoji
Smileys	Any smiley e.g. :)
Posessive pronouns - second person	yours

Table 14 – lexicon of features

**Indicates all variations of words mentioned were used*

Conflicts of Interest

Miss Isabel Holmes
None

Dr Nelli Ferenczi
None

Dr Timothy Cribbin
None