Supplementary Material for:
Down the Rabbit Hole: Detecting Online Extremism, Radicalisation, and Politicised Hate Speech

JAROD GOVERS, ORKA Lab, Department of Software Engineering, University of Waikato, NZ
PHILIP FELDMAN and AARON DANT, ASRC Federal, US
PANOS PATROS, ORKA Lab, Department of Software Engineering, University of Waikato, NZ

1 DEFINITIONS—THE ALGORITHM HANDBOOK

This document includes the supplementary material referenced in the main *Down the Rabbit Hole: Detecting Online Extremism, Radicalisation, and Politicised Hate Speech* **Systematic Literature Review (SLR)**. This document offers a "dictionary/look-up table" for the core algorithmic architectures for the non-deep machine learning and deep learning models mentioned throughout the SLR, alongside other findings and design considerations. We contextualise the relevant strengths and weaknesses of the various algorithmic approaches for text and visual models for **Extremism, Radicalisation, and Hate Speech (ERH)** detection. No new findings are in this handbook/"look-up table." Hence, those familiar with the models listed in the contents above need not read this section.

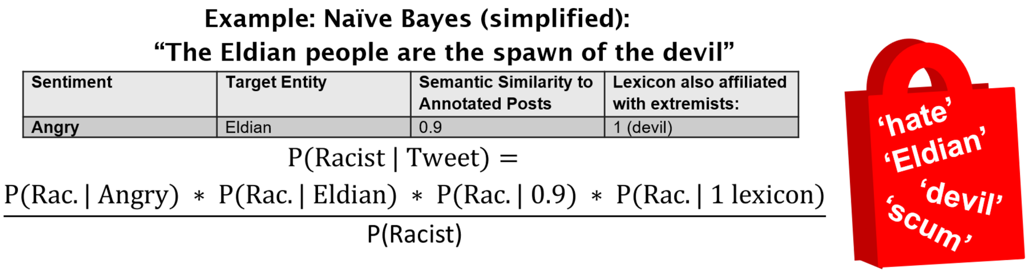
1.1 Definitions for Traditional (Non-deep) Machine Learning Algorithms

We aggregate common and historic non-deep machine learning algorithms into the "traditional" **Machine Learning Algorithm (MLA)** category. Hence, this section defines each of the baseline models used for textual or community detection models—consisting of:

(1) Sentimental Bag of Words approaches,
(2) Naïve Bayes,
(3) Decision Trees,
(4) Support Vector Machines,
(5) Clustering Models.

1.1.1 Bag of Words (BoW). **Bag-of-Words (BoW)** approaches simplify complex contextual sentences into a multiset ("bag") of individual words by assigning a value or probability to each word in its relation to a specific document class. For instance, a BoW approach would deconstruct the contiguous sentence, "The Eldian people are the spawn of the devil" (where Eldian is a fictitious race), into an unordered bag of individual words. While "are" and "the" are unlikely to have a considerable influence on whether a sentence is hate speech or not, the use of "devil" and "Eldian [race]" is more frequently paired in hate speech than for non-hateful/off-topic text. The disregard of word order and the relationship of BoW approaches, and MLA models at large, constitute context-insensitive models. For instance, a BoW model does not know that "I love the Eldian people but hate their food" is paring love -> Eldian, and hate -> food, and thus would consider "I hate

0360-0300/2023/07-ART319 $15.00
https://doi.org/10.1145/3583067



Example: Naïve Bayes (simplified):
"The Eldian people are the spawn of the devil"

Sentiment	Target Entity	Semantic Similarity to Annotated Posts	Lexicon also affiliated with extremists:
Angry	Eldian	0.9	1 (devil)

$$P(\text{Racist} \mid \text{Tweet}) = \frac{P(\text{Rac.} \mid \text{Angry}) * P(\text{Rac.} \mid \text{Eldian}) * P(\text{Rac.} \mid 0.9) * P(\text{Rac.} \mid 1 \text{ lexicon})}{P(\text{Racist})}$$

Fig. 1. An abstracted example of Bag-of-Words approach within a Naïve Bayes classifier—demonstrating its lack of context sensitivity and the focus on key *racist* words for ERH detection tasks.

the Eldian people but love their food" as identical. Likewise, BoW approaches do not consider alternate word meanings/uses (e.g., "I ran for government" versus "I ran away"). Nonetheless, BoW approaches are core to word-specific 'blacklists' in content moderation, such as banning users who use slurs in a post. However, for nuanced and often politicised discussions on controversial topics, simple blacklists can lead to injurious censorship—due to the context and use of such words.

Sentimental algorithms, such as SentiStrength [42] aggregate individual words into individual emotions—whereby "love" indicates a positive sentiment, while "hate" generally appears in vitriolic speech. Figure 1 outlines an abstracted representation of the sentiment classification based on the average sentiment score of a sentence. However, the context-insensitive BoW models again fails for nuanced cases, whereby Sharma et al. [41] identified that SentiStrength cannot detect negations (e.g., "I am NOT happy" where happy skews the final sentiment scores).

1.1.2 Naïve Bayes. Naïve Bayes classifiers represent types of probabilistic classifiers utilising Bayes theorem with the assumption that the influence of each variable for classification is independent of each other (i.e., *naïve*) [44]. For document classification, notable features are assigned a probability for their occurrence given a specific class. For instance, a hate speech post that has an angry sentiment may have a P(0.8) (Probability of 80%) of being hateful, given that a test hate speech dataset may be 80% angry speech. Bayes rule represents these chains of (assumed) independent/unrelated probabilities to form a final probability for a test instance.

Notable features for probability models include:

- *Textual features*—(e.g., sentimental scores, appearance of certain slurs/terms),
- *Network data* (e.g., probability that someone who is friends with a supremacist is also a supremacist, retweet relationships),
- *Metadata* (e.g., length of a post, readability via a Flesch Reading Ease score, number of posts).

In the example of Figure 1, the probability that the tweet is racist depends on the probability that the racist tweet is angry, contains racial terms ("Eldian"), the semantic similarity between known hate speech posts, and the appearance of a negative lexicon. Naïve Bayes can be a final classifier for aggregating context-sensitive embeddings (e.g., deep learning models) and multiple "ensembles" of approaches/models—via chaining their probabilities together with this Bayes rule.

1.1.3 Decision Trees. Chaining the correlations between features and their class likelihood can also span a tree of scenarios. If an annotated dataset indicates that a post is 80% likely to be racist if a sentiment-scoring algorithm detects anger, then a binary decision emerges—if post contains angry words, then likely hate speech; if not, then not hate speech. These rules construct decision trees, where the root constitutes the instance (text, network, metadata, or image), and each node is a decision, with the leaves (final node) being the expected class value (i.e., the classification) [44].

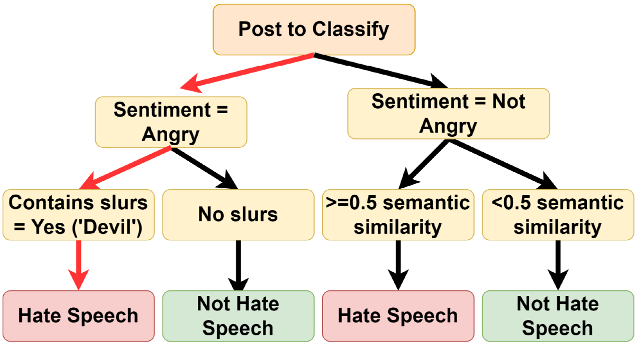


Fig. 2. An example of a decision tree, with the leaf nodes constituting the classification. In the Eldian hate speech example, this would require traversing the left branches recursively for the final "Hate Speech" classification leaf (shown via the red arrows).

Hence, decision trees are not naïve as they rely on specific values of other features when traversing a tree's branches for a prediction.

Creating an optimal tree that maximises accuracy and precision is not trivial due to the feature explosion of possible rules and tree nodes. Hence, Random Forest classifiers rely on a divide-and-conquer algorithm for generalising feature pairings into class classifications with a random initialisation [18]. This recursive process requires finding optimal splits to maximise the separation of classes for a final leaf, with an example tree presented in Figure 2—where a random forest would consists of multiple trees as a *forest*. Ideally, a leaf node should encapsulate instances of one class.

Random forests generate multiple decision trees and select the final prediction based on the predictions from the majority of decision trees. Utilising multiple trees with a random initial tree state increases the range of features and values selected during the training step. Utilising multiple trees and testing the models on untrained "test" data minimises the risk of over-fitting to the training (i.e., a classifier that performs reliably on the training dataset but not on real-world data).

Random forests strengths include its ability to tie dependent and complex features while reducing over-fitting through pruning (i.e., reducing tree size to generalise the model). Hence, decision trees capture related concepts in hate speech where naïve BoW approaches do not—such as the appearance of anger/negative sentiment invoking the use of charged terms (e.g., racism as an emotional outlet) or frequency of posts and sentiment.

1.1.4 Support Vector Machines (SVM). SVMs are another supervised learning model for classification and regression tasks, seeking to map instances in vector spaces to maximise the distance between classes [14], visualised in Figure 3. Mapping features to multidimensional vectors can exponentially increase dimensions (an issue shared in deep-learning models). Thus, SVMs reduce irrelevant features through specific kernels—typically a linear, polynomial, Gaussian or sigmoid function. These kernels reduce the feature set to draw boundaries between two classes, similar to logistic regression. These boundaries are either hard (i.e., a binary classification) or soft—allowing outliers near the boundary for edge cases, like niche controversial and offensive, but not ostensibly targeting protected characteristics. SVM models are computationally faster and reduce memory compared to deep learning models [3, 45], while achieving comparative performance outlined in RQ4. Dimensionality reduction techniques can also reduce runtime by reducing the complexity of large feature spaces from textual or network data, such as via Principle Component Analysis [27].

SVMs are the consistently highest performing MLAs per RQ4, while lowest complexity, with $O(m * n)$ complexity for a Linear Kernel SVC—where m = feature count, and n = number of instances.

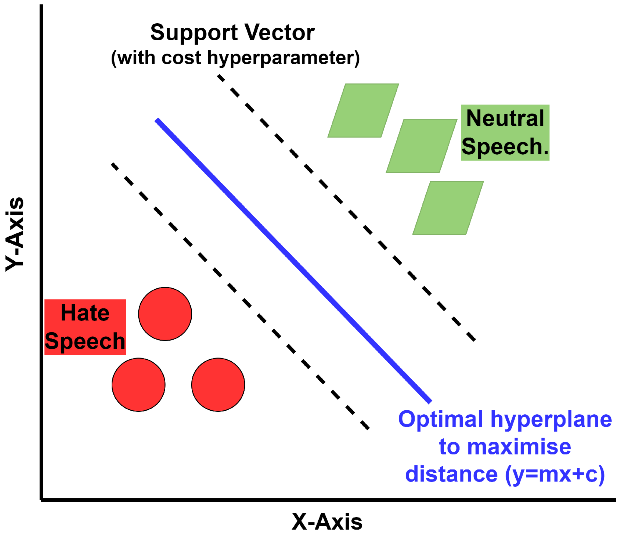

Support Vector
(with cost hyperparameter)
Neutral Speech.
Y-Axis
Hate Speech
Optimal hyperplane to maximise distance (y=mx+c)
X-Axis

Fig. 3. Support Vector Machine where instances beyond the boundaries (support vectors) are automatically assigned to the class.

1.1.5 Clustering and Nearest Neighbour Classifiers. Instead of annotated hate speech datasets, clustering methods group by textual similarity via ***Natural Language Processing*** **(NLP)**, and network relations via *Community detection*. Hence, clustering can work in cases of fully annotated datasets as supervised learning, semi-annotated datasets as semi-supervised learning, or unlabelled raw web scrapped data for unsupervised learning.

For supervised learning, ***K-Nearest Neighbour*** **(KNN)** classifiers work via evaluating the nearest neighbours' likeliness when projecting the textual, network, or metadata features onto a multidimensional space [2]. The "distance" between feature spaces typically rely on Euclidean, Manhattan, or Minkowski distance—where the latter two are suited for non-linear feature spaces. Non-euclidean distances are ideal where dimensions are not comparable, as Manhatten distance reduces noise/errors from outliers, since the gradient has a constant magnitude.

Clustering examples for hate speech detection includes K-Means, which partitions n observations into k clusters [27]. K-Means automatically generates clusters, thus does not require annotated datasets. Hence, K-Means can detect novel groups, including emergent extremist organisations, or influential individuals [4]. Unsupervised clustering's strength for ERH detection is how it circumvents the definition issues for annotating data and can cluster large movements without costly annotation. However, K-Means may not identify manifestly hateful posts, as it does not abide by any standard imbued within strict annotation criteria. Evidently, in the cross examination of a naïve approach versus their proposed K-Means derived model by Moussaouri et al. [30], the naïve approach outperformed the possibilistic clustering by 0.07–0.14 for accuracy 0.04–0.05 for precision.

1.2 Definitions for Deep Learning Approaches

Deep learning represents a family of machine learning algorithms with multiple layers and complexity, typically via neural network architectures. Neural networks rely on training a network with a set of weights at each layer, known as *neurons*. The first layer of a neural network utilises numeric representation of an instance (e.g., hateful text) in numeric "tokenised" form, which is adjusted throughout the hidden lower layers towards a final output (typically) classification

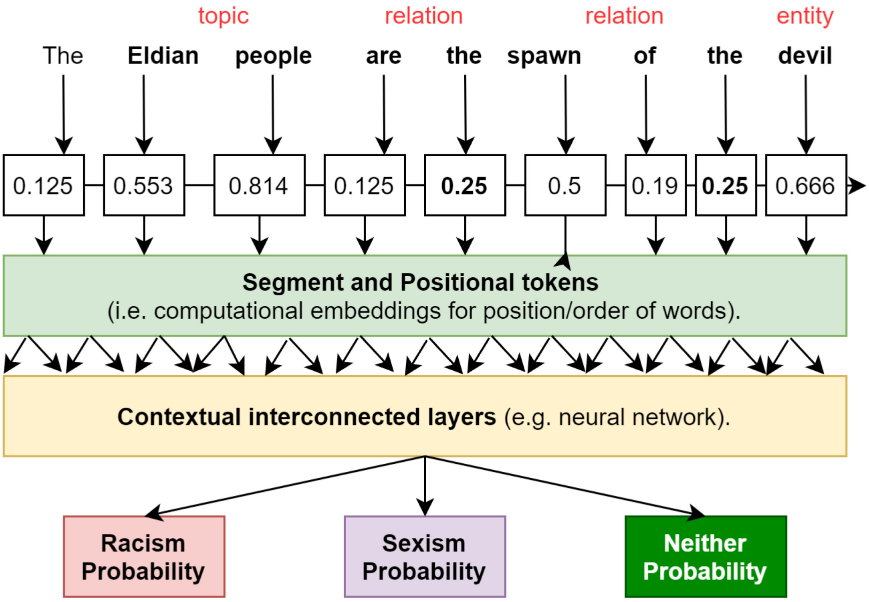


Fig. 4. An abstracted example of a neural network for text. The top text represents its raw syntactic form, with its converted numeric embedding representation. These embeddings are responsible for altering the weights to increase token prediction or generation (for transformers) via backpropagation. The final output layer for this example would offer the probability that the given text is racist, sexist, or benign.

layer. Each downwards training step results in readjusting the weights of the upper layers for the neurons—known as *backpropagation* [38]. Figure 4 displays this architecture for neural networks per our example. The benefit of **Deep Learning Algorithms (DLAs)** in ERH detection is the preservation of word order and meaning (e.g., “I ran” versus “I ran for president”), thus displaying context-*sensitivity*. Given dual-use words such as “queer” or racially motivated slurs, understanding the surrounding contextual words is essential to reduce bias via misclassifications [31]. DLAs dominate the benchmark dataset leader-board in RQ4.

1.2.1 Convolutional Neural Networks (CNN). **Convolutional Neural Networks (CNNs)** expand on the neural network model through a convolutional layer—which acts as a learnable filter for textual or image embeddings [44]. Moreover, CNNs include a pooling layer(s) to reduce the spatial complexity of the network’s features. Reducing spatial size helps reduce the number of parameters and thus training time and memory footprint, while reducing over-fitting by generalising patterns in the training data.

1.2.2 Long Short-term Memory (LSTM) and Gated Recurrent Unit (GRU). **Long Short-term Memory (LSTM)** and **Gated Recurrent Unit (GRU)** aim to increase contextual awareness to process data sequences with long-term gradients to retain information on prior tokens [12, 19]. LSTM and GRU seek to reduce the vanishing gradients caused during backpropagation steps, which reduces classification performance as older trained instances are effectively “forgotten” due to later weight changes. Similarly, GRU’s are a gating approach with fewer parameters and thus higher runtime, enabling larger neural networks overall. CNN models with LSTM and GRU connections outperform CNNs on their own for hate speech detection [22, 23, 32]. The highest performing BiLSTM model expands LSTM for bidirectional input, via two LSTMs—where tokens in the network utilise information from past (backwards) tokens/data and future (forwards) data [32]. The ability to uphold the temporal memory of prior tokens (attention) constitutes a **Recurrent Neural Network (RNN)**.

1.3 Language Transformer Models

The state-of-the-art transformer architecture relies on *self-attention*—the memory retention of neural networks where each token of a sequence is differentially weighted [8, 16]. Unlike Recurrent Neural Networks (i.e., neural networks where nodes follow a temporal sequence), a transformer's attention mechanism utilises context for any position for the token sequence. Hence, transformers can handle words out of order to increase understanding. Transformers offer greater classification performance (see RQ4) at the expense of memory and computational overhead. A considerable ethical threat of transformer models is their capability to predict future tokens (i.e., text generation). For instance, a malicious actor could create realistic automated trolls or radicalising synthetic agents as bots. Language models also risk data leakage of their trained data through predicting tokens found in the original trained dataset, such as names or addresses [8].

1.3.1 Cross-encoders (e.g., BERT). **Bidirectional Encoding Representations from Transformers (BERT)** is the most common cross-encoder observed for ERH detection [16], with the highest performance of all NLP models. Cross-encoders offer higher performance for classification tasks, through retaining information over a given sequence with a label (i.e., self-attention). BERT's core strength is its memory retention of all tokens in a sentence, thus upholding full context-sensitivity of every word in the post it seeks to classify. However, cross-encoders are computationally expensive due to high parameter counts (110 million parameters for BERT-base, 365 million for BERT-large), an issue further outlined in RQ4. Hence, an area of ongoing research includes model distillation (optimising and reducing parameter count to reduce memory requirements and training time), specialised training datasets, and alternate layers [26, 37, 46]. BERT is pre-trained on entries from English Wikipedia (2.5 million words) and the English BookCorpus (800 million words) [16]. Hence, such pre-trained models are then fine-tuned on a smaller dataset (typically 1,000+ instances, per RQ2's benchmark datasets) to optimise the BERT weights to detect hate speech with the context of its pre-trained corpus.

1.3.2 Generative Pre-trained Models (GPT). Similarly, the state-of-the-art GPT transformer architecture expands on the encoder blocks (shared with BERT) to include decoder blocks [8]. Hence, GPT works on a single token (i.e., word vector) and produces estimates for the sequence's next token—ideal for tasks such as text generation, summarising, question answering, and information retrieval.

GPT models differ from BERT-based models via *masked self-attention*—an alternate form of context-sensitivity where the model only knows the context of the prior words in the sentence. GPT-2/3 [8], GPT-Neo [7], and Jurassic-1 [25], are notable 2019-2021 era multi-billion parameter models—where larger datasets and parameter count result in more human-like text generation and higher performance in information retrieval tasks [8].

GPT's core strength in ERH detection synthetic hate speech generation via a GPT model fine-tuned on a hateful corpus—as investigated by Wullah et al. (see RQ3) [45]. However, state-of-the-art GPT models utilise up to 178B parameters, whereby memory and computational requirements scale linearly. Hence, future GPT work in synthetic text generation should consider inference tasks over fine-tuning. Specifically, inference utilises a pre-trained model's on-demand text generation capability through prompts rather than altering each of the billions of weights. Using the auto-complete-like inference capabilities for generating realistic synthetic hate speech posts constitutes a novel case of prompt engineering in ERH detection and thus is a potential future research project.

1.4 Definitions for Prominent Feature Extraction Techniques

This subsection outlines the three most common feature extraction techniques used for textual ERH detection—as outlined in RQ2 in the SLR. These models seek to identify hateful lexicons from text, or create numerical representations for word or sentence meaning via embeddings. We deconstruct the six most common feature extraction techniques observed in our SLR.

1.4.1 Word2Vec. Word2Vec is a model to convert words into vector embeddings, which compares synonymous words (e.g., "hate" and "disgust") via numerical vectors [29]. Word2Vec compares these word-to-vector embeddings via *semantic similarity* by evaluating their cosine similarity between their vectors (e.g., comparing word vectors of an unknown class instance to words from a known "hate speech" instance(s) to make a "hate or not" classification). On a word-level basis, the vector value for "king" – value for "man" + value for *woman* would result in a vector similar to *queen* [29]. In our case, a "Islamist extremist" and "ISIS" are semantically similar akin to "White Supremacy" and "Nazism."

1.4.2 Doc2Vec. Similar to Word2Vec, Doc2Vec aggregates vector embeddings for *paragraphs* in addition to individual words [24], thereby offering memory of the current context and paragraph's topic—useful for understanding a whole post's sentiment and meaning.

1.4.3 N-grams. N-grams represent contiguous sequences of n-number of characters for frequency analysis given their non-linear distribution in English, as well as when comparing a radical versus non-radical corpus [44]. This linguistic model is often paired with methods such as TF-IDF or BoW.

1.4.4 Term Frequency-Inverse Document Frequency (TF-IDF). **Term Frequency-Inverse Document Frequency (TF-IDF)** determines the relevance of a word in a document by comparing its frequency *in the document* compared to its inverse number for the frequency of that word *across all documents* [44]. Thereby, assigning each word a weight to signify its semantic importance compared to the wider corpus. For instance, radical Islamist *dog-whistle terms* (i.e., coded or suggestive political messages intended to support a group) appeared disproportionately in extremist text compared to a neutral religious corpus [36].

1.4.5 SenticNet. SenticNet embeds pattern matching, parser trees, and LSTM-CNN models for sentiment analysis, with the aim to replace a naïve BoW approach within a proclaimed bag of concepts and narratives [10]. Specifically, it includes feature extraction methods of concept parsing (i.e., understanding linguistic patterns in natural language into conceptual pairs), subjectivity and polarity inference, alongside personality and emotion extraction.

1.4.6 Global Vectors for Word Representation (GloVe). **Global Vectors for Word Representation (GloVe)** offers an unsupervised learning algorithm for context-independent word-to-vector embeddings [34]. While similar in creating vectors akin to Word2Vec, GloVe instead establishes word co-occurrences using matrix factorization (i.e., co-occurrence matrix of word [row] and context [usage of the word in the document]) and dimensionality reduction techniques.

2 SLR DESIGN CONSIDERATIONS

This supplementary material section outlines the additional criteria and considerations for selecting papers and ensuring privacy-protections for users, groups and collected data. In essence, this section offers a meta-analysis of the ethics and selection process used throughout the SLR.

2.1 Quality Assessment Criteria

The following includes our paper inclusion quality check criteria—with a score of 13 or higher required for inclusion in the final paper selection (i.e., final 51 papers included).

We propose a critical criteria for quality assessment to filter irrelevant or ambiguous studies. Specifically, for a study that passed a title and abstract screen, we assess the study's clarity for ERH definitions and annotations (for objective and legible classifications), methodical clarity (i.e., outlining each study's algorithmic model, methods, data collection processes, and statistical analysis/evaluation methods), and socio-technical considerations. We weighted each quality assessment section to prioritise their research methodology and clarity in their technical methods over their *Conceptual Quality* for studies encompassing broader socio-technical issues such as ethics, legality, or ERH clarity. After a ten-paper pilot study, we selected a score threshold of 65% to exclude irrelevant or ambiguous studies. Our supplementary material document includes the criteria and scoring for our quality assessment rubric.

2.1.1 Computational Quality (0 = None, 1 = Partial, 2 = Full).

(1) Is the radicalisation/affiliation detection model clearly defined?
(2) Is the radicalisation/affiliation detection model's algorithm clearly defined?
(3) Is the training data reputable?
(4) Are the models results compared to similar state-of-the-art methods?
(5) Is the methodology for designing and conducting their experiment clearly defined?
(6) Are patterns and trends discussed and presented clearly?

2.1.2 Epistemological Quality (0 = None, 1 = Partial, 2 = Full).

(1) Does the source(s) (data or researchers) avoid any conflict of interests or expressed biases? (i.e., explicit support/funding from a political think tank or state agency).
(2) Does the study provide a cited or evidence-based definition for "radicalisation," "hate speech," or "extremist" affiliation?
(3) Are the dataset annotations vetted by more than one annotator to reduce bias?

2.1.3 Conceptual Quality (0 or 0.5 Value, as not Critical but Useful).

(1) Does the study discuss social or ethical issues in ERH detection (e.g., censorship)?
(2) Do the authors discuss the legality of their model or definitions?
(3) Does the study evaluate its model across multiple social media platforms?
(4) Does the study discuss regulatory frameworks or recommendations for social media platforms based on their findings?

2.1.4 Researcher Ethics. We focus on key terms and compositions of ERH examples to protect the privacy of the individuals exposed, as recommended by meta-studies on extremism research ethics [9, 13, 28]. When linking ERH detection to real-world groups and events, we solely focus on events and organisations that resulted in media attention or criminal convictions. In no part during this SLR did we attempt to track users, groups, or correlate online users to any personally identifiable information (name, location, username, etc.) given the ease of composing online data into a traceable online fingerprint.

Similar to the social norms in New Zealand in the aftermath of the Christchurch shooting, no extremists, terrorists, and/or criminals are referred by name to minimise publicity. We recognise the potential for political or cultural bias in this charged field by citing international non-partisan Non-governmental Organisations when framing ERH concepts, and avoid searching any party or ideology in our search strategy. Moreover, we encourage that our findings and recommendations invoke an open debate among social media platforms, governments, and the wider public. However, we do not condone the use of ERH detection in social media as a form of autonomous law. We recommend human-in-the-loop processes when handling or classifying data via independent reviews, privacy protections, and complaint and redress mechanisms for deployed models.

Our recommendations thereby focus on **Open Source Intelligence (OSINT)**-oriented studies that do not consider governmental or private-conversation surveillance (with the exception of one hybridised study that appeared in our search [20]). We thereby consider ERH detection as a *computational* method aimed at garnering community-insights, trends, and flagging for *social media platforms themselves* to use. Whether ERH detection *policies* should encourage deplatforming, deranking, demonetisation, fact-checking, or targeted counter-speech/prevention programs require further research. We encourage open interdisciplinary research in public and private-communications—particularly ethical and legal discussions.

3 THE CASE FOR PERFORMANCE ENGINEERING WHEN EVALUATING MODELS

While high F1-scores help enforce community guidelines via accurate predictions and reduce injurious censorship from false positives, runtime performance trade-offs are seldom discussed. DLAs may perform within 1% (F1-score) of their MLA counterparts in NLP studies but require significantly higher computational resources. For instance, fine-tuning a BERT-large model for NLP tasks requires **Graphics** or **Tensor Processing Units** (**GPU** or **TPU**), restricting researchers from testing large language models [45, 47]. For community detection, uncompressed network models can include up to 27.4 million links [6], which significantly increases computational and memory requirements for a minimal 1–5% performance gain. Specifically, using a **Possibilistic Approach (PA)** with dimensionality reduction reduced subgraph mining runtime by up to 67% (1,500 to 500 s on an 8-core 3.2 GHz system), while reducing accuracy by only 4% [30]. Furthermore, *community-level insights* on topics with millions of tweets, relations, and discussions can lead to a network *explosion* with a non-deterministic polynomial runtime [5, 30]. In graph-detection approaches, performance engineering and optimisation for mining frequent subgraphs and graph-traversal is an active area of research [30]). No NLP studies consider performance engineering for DLAs despite developments in model distillation and sentence-level embeddings [37].

Thus, we recommend that researchers consider performance trade-offs in future work and investigate a possible standardised performance-complexity metric (e.g., parameter count versus F1-score ratio) to build scalable, energy-efficient and fiscally viable models. Moreover, fine-tuning or retraining DLAs, or regenerating frequent subgraphs for community detection, should be a frequent endeavour to adapt to the rapidly evolving topics, entities, and events throughout online discourse. Due to the computational costs of fine-tuning or training multi-billion parameter models, we recommend approaches that do not require expensive training, such as few-shot learning (i.e., giving several known instances of ERH and a unseen "test" instance) and prompt engineering [8].

4 UPTAKE ROADMAP EXPANDED

This supplementary section expands on the dataset and model research gaps highlighted in Figure 16 of the main *Down the Rabbit Hole* SLR document. We categorise these research recommendations into eight core components for our proposed *ERH Context Mining* research field, which we aggregate and visualise in Figure 5.

4.1 Model Recommendations

The two predominant recommendations for future work are investigating the role of *changes in hateful affiliation or speech over time* to satisfy the temporal requirement for *Radicalisation*

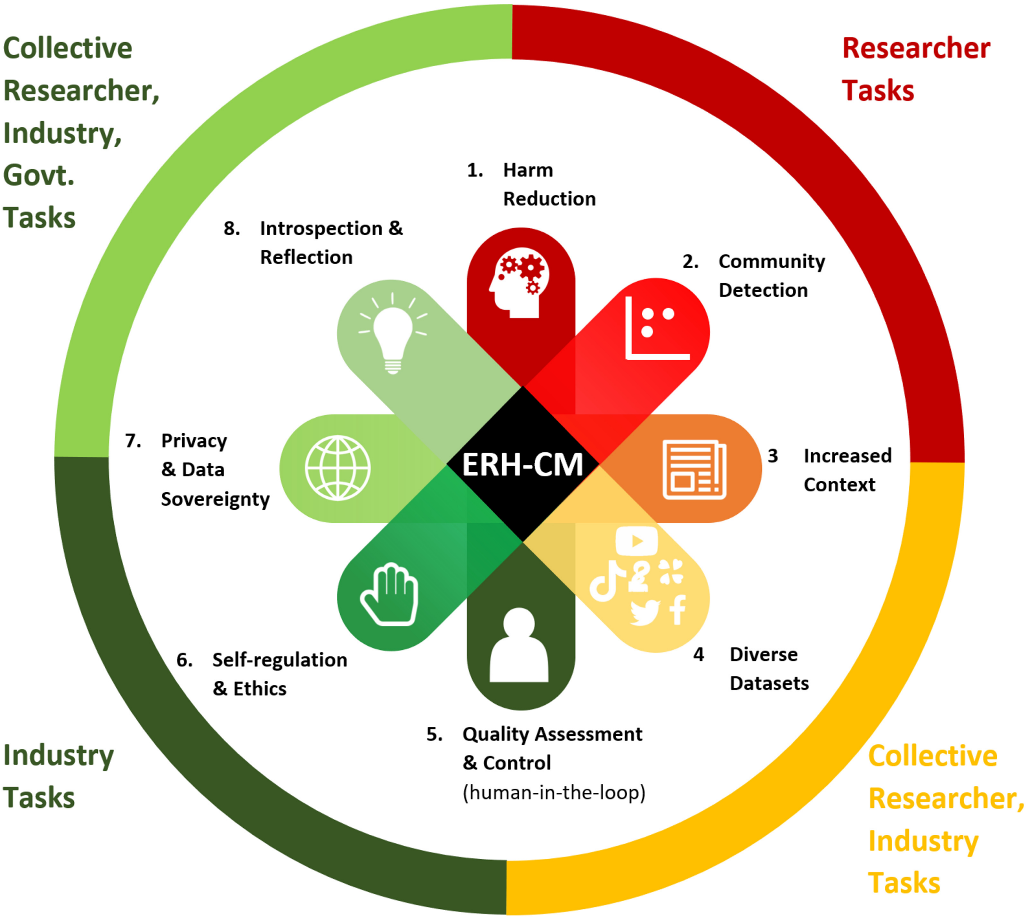

Collective Researcher, Industry, Govt. Tasks
Researcher Tasks
Industry Tasks
Collective Researcher, Industry Tasks
ERH-CM
1. Harm Reduction
2. Community Detection
3 Increased Context
4 Diverse Datasets
5. Quality Assessment & Control (human-in-the-loop)
6. Self-regulation & Ethics
7. Privacy & Data Sovereignty
8. Introspection & Reflection

Fig. 5. ERH Context Mining (ERH-CM) eight core components for Research, Industry, and Government.

detection, and to train models on multiclass datasets from multiple platforms. We note that only one study considered temporal data on both meso and macro (changes within and between groups), and micro (individual) levels, although recommended as future work within four other studies [3, 11, 20, 40]. Moreover, we recommend expanding on DLAs as the target for future research based on their leading performance in RQ4. Neural language models offer a macro-level societal understanding due to their pre-trained corpus on academic sources, OpenWebText2 Reddit discussions, and Wikipedia [8]. Furthermore, transformer models beyond 764 million parameters are untested.

Bot, troll, meme, entity, dis/misinformation, and satire detection remain underdeveloped—which could lead to censorship or undermine democratic institutions. Five studies recommended multimedia detection as future work [1, 5, 11, 17, 35].

To protect user privacy from recreating user content from neural language models, we encourage *privacy-by-design* software engineering through machine learning paradigms such as **Differential Privacy (DP)**. DP-paradigm models and datasets reduce the potential for self-identification from trained models (i.e., data leakage, such as names or usernames in open-source datasets), as DP-paradigm models use pseudo-anonymised *patterns* of groups and hate.

4.2 Dataset Recommendations

To investigate the roles of radicalisation, we recommend expanding on the dataset annotation approach by de Gibert et al. [15] by creating a *conversation-level* dataset with public non-hateful replies to a post for context. Moreover, future benchmark datasets should consider pulling data across platforms to investigate macro-level radicalisation trends between platforms. We note that only two studies considered anti-Asian sentiment in COVID-related tweets, targeting a seldom explored topic and demographic [21, 39] worthy of expansion given the ongoing COVID-19 pandemic.

Likewise, future datasets should consider the role of indigenous discussions and potential researcher biases given the Anglo-dominant field of ERH research. Given the rise of COVID extremism [43], far-right movements, and xenophobia in Oceania. Hence, we recommend geotargeted datasets to consider the differences for investigating ERH topics, which would demonstrate NZ's commitment to our Christchurch Call to Action Summit. Investigating unexplored and minority groups could also provide imperative insights for social scientists regarding the conversational dynamics, morphological mapping, and ideological isomorphism from radical minority groups towards the majority. Likewise, research on vulnerable communities (youth, gender and sexual minorities, religious, racial, and geographically distant peoples) would aid social media platforms in both identifying unique radicalising risks, as well as avenues for support and de-escalation. In the mental health end, we recommend building on Nouh et al.'s proposed approach of extracting textual, psychological and behavioural features [33], both due to its performance, as well as its potential for analysing societal *factors and ERH roots* such as correlations between mental health issues (isolation, depression, etc.) and vulnerability to radicalisation towards violent extremism.

For any counter-extremism or de-radicalisation studies, we recommend work in ethical and legal guidelines to protect privacy, avoid backlash or inadvertent algorithmic amplification.

Investigating posts from periods of political, or social crisis (e.g., COVID health measures, post-terror attack discourse, etc.) could also help identify cases of ERH on mainstream platforms before they are deplatformed/removed. Event-based datasets would provide unique sociological insights on the role of societal stress and emergencies on the human psyche and online group dynamics.

To reduce the cost, variability in inter-annotator agreement, and psychological impact of human annotation, we recommend unsupervised clustering-based research and propose using synthetic conversational agents to simulate extremist discourse. Simulating online radicalisation in a closed environment would present a safe, ethical, and non-invasive method to build benchmark datasets.

REFERENCES

[1] Umar Abubakar, Sulaimon A. Bashir, Muhammad B. Abdullah, and Olawale Surajudeen Adebayo. 2019. Comparative study of various machine learning algorithms for tweet classification. *i-manager's J. Comput. Sci.* 6 (01 2019), 12. https://doi.org/10.26634/jcom.6.4.15722

[2] Naomi S. Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *Amer. Stat.* 46, 3 (1992), 175–185. https://doi.org/10.1080/00031305.1992.10475879

[3] Oscar Araque and Carlos A. Iglesias. 2020. An approach for radicalization detection based on emotion signals and semantic similarity. *IEEE Access* 8 (2020), 17877–17891. https://doi.org/10.1109/ACCESS.2020.2967219

[4] Alon Bartal and Gilad Ravid. 2020. Member behavior in dynamic online communities: Role affiliation frequency model. *IEEE Trans. Knowl. Data Eng.* 32, 9 (2020), 1773–1784. https://doi.org/10.1109/TKDE.2019.2911067

[5] Matthew Benigni, Kenneth Joseph, and Kathleen M. Carley. 2018. Mining online communities to inform strategic messaging: Practical methods to identify community-level insights. *Comput. Math. Org. Theory* 24, 2 (2018), 224–242.

[6] Matthew C. Benigni, Kenneth Joseph, and Kathleen M. Carley. 2017. Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter. *PLoS ONE* 12, 12 (2017), 23.

[7] Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. Retrieved from https://zenodo.org/record/5297715#.Y-ujk3bMKUk. https://doi.org/10.5281/zenodo.5297715

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., Vancouver, Canada, 1877–1901. Retrieved from https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[9] Elizabeth Buchanan. 2017. Considering the ethics of big data research: A case of Twitter and ISIS/ISIL. *PLoS ONE* 12, 12 (Dec. 2017), 1–6. https://doi.org/10.1371/journal.pone.0187155

[10] Erik Cambria, Soujanya Poria, Devamanyu Hazarika, and Kenneth Kwok. 2018. SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. *Proc. AAAI Conf. Artific. Intell.* 32, 1 (Apr. 2018), 1795–1802. https://ojs.aaai.org/index.php/AAAI/article/view/11559.

[11] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. *The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data.* Association for Computing Machinery, New York, NY, 3175–3187.

[12] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Retrieved from https://arxiv.org/pdf/1406.1078.pdf.

[13] Maura Conway. 2021. Online extremism and terrorism research ethics: Researcher safety, informed consent, and the need for tailored guidelines. *Terror. Politic. Viol.* 33, 2 (2021), 367–380. https://doi.org/10.1080/09546553.2021.1880235

[14] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Mach. Learn.* 20, 3 (Sep. 1995), 273–297. https://doi.org/10.1023/A:1022627411411

[15] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW'18).* Association for Computational Linguistics, 11–20. https://doi.org/10.18653/v1/W18-5102

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[17] Margeret Hall, Michael Logan, Gina S. Ligon, and Douglas C. Derrick. 2020. Do machines replicate humans? Toward a unified understanding of radicalizing content on the open social web. *Policy Internet* 12, 1 (2020), 109–138.

[18] Tin Kam Ho. 1995. Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Vol. 1. IEEE Press, 278–282 vol. 1. https://doi.org/10.1109/ICDAR.1995.598994

[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9 (12 1997), 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735

[20] Benjamin W. K. Hung, Anura P. Jayasumana, and Vidarshana W. Bandara. 2016. Detecting radicalization trajectories using graph pattern matching algorithms. In *Proceedings of the IEEE Conference on Intelligence and Security Informatics (ISI'16).* IEEE Press, 313–315. https://doi.org/10.1109/ISI.2016.7745498

[21] Bokang Jia, Domnica Dzitac, Samridha Shrestha, Komiljon Turdaliev, and Nurgazy Seidaliev. 2021. An ensemble machine learning approach to understanding the effect of a global pandemic on twitter users' attitudes. *Int. J. Comput., Commun. Control* 16, 2 (2021), 11. https://doi.org/10.15837/ijccc.2021.2.4207

[22] Prashant Kapil and Asif Ekbal. 2020. A deep neural network based multi-task learning approach to hate speech detection. *Knowl.-Based Syst.* 210 (2020), 106458. https://doi.org/10.1016/j.knosys.2020.106458

[23] Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the 1st Workshop on Trolling, Aggression and Cyberbullying (TRAC'18).* Association for Computational Linguistics, 1–11. Retrieved from https://aclanthology.org/W18-4401.

[24] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning (ICML'14)*, Vol. 32. JMLR, 1188–1196.

[25] Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. 2021. *Jurassic-1: Technical Details and Evaluation.* Technical Report. AI21 Labs.

[26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Retrieved from https://arxiv.org/abs/1907.11692.

[27] Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE Trans. Info. Theory* 28, 2 (1982), 129–137. https://doi.org/10.1109/TIT.1982.1056489

[28] Alice E. Marwick, Lindsay Blackwell, and Katherine Lo. 2016. Best practices for conducting risky research and protecting yourself from online harassment (Data & Society Guide). Retrieved from https://datasociety.net/pubs/res/Best_Practices_for_Conducting_Risky_Research-Oct-2016.pdf.

[29] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. Retrieved from https://arxiv.org/abs/1301.3781.

[30] Mohamed Moussaoui, Montaceur Zaghdoud, and Jalel Akaichi. 2019. A possibilistic framework for the detection of terrorism-related Twitter communities in social media. *Concurr. Comput.: Pract. Exper.* 31, 13 (2019), 20. https://doi.org/10.1002/cpe.5077

[31] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PLoS ONE* 15, 8 (Aug. 2020), 1–26. https://doi.org/10.1371/journal.pone.0237861

[32] Usman Naseem, Imran Razzak, and Ibrahim A. Hameed. 2019. Deep context-aware embedding for abusive and hate speech detection on Twitter. *Austral. J. Intell. Info. Process. Syst.* 15, 3 (2019), 69–76.

[33] Mariam Nouh, Jason R. C. Nurse, and Michael Goldsmith. 2019. Understanding the radical mind: Identifying signals to detect extremist content on Twitter. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI'19)*. IEEE Press, 98–103. https://doi.org/10.1109/ISI.2019.8823548

[34] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. Association for Computational Linguistics, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[35] Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: Political ideology prediction of Twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 729–740. https://doi.org/10.18653/v1/P17-1068

[36] Zia Ul Rehman, Sagheer Abbas, Muhammad Adnan Khan, Ghulam Mustafa, Hira Fayyaz, Muhammad Hanif, and Muhammad Anwar Saeed. 2021. Understanding the language of ISIS: An empirical approach to detect radical content on Twitter using machine learning. *Comput. Mater. Continua* 66, 2 (2021), 1075–1090.

[37] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. Association for Computational Linguistics, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[38] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1985. *Learning Internal Representations by Error Propagation*. Technical Report. University of California San Diego, La Jolla Institute for Cognitive Science.

[39] Furqan Rustam, Madiha Khalid, Waqar Aslam, Vaibhav Rupapara, Arif Mehmood, and Gyu Sang Choi. 2021. A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *PLoS ONE* 16, 2 (Feb. 2021), 1–23. https://doi.org/10.1371/journal.pone.0245909

[40] Ryan Scrivens, Garth Davies, and Richard Frank. 2018. Searching for signs of extremism on the web: An introduction to Sentiment-based Identification of Radical Authors. *Behav. Sci. Terror. Politic. Aggress.* 10, 1 (2018), 39–59. https://doi.org/10.1080/19434472.2016.1276612

[41] Ankur Sharma, Navreet Kaur, Anirban Sen, and Aaditeshwar Seth. 2020. Ideology detection in the indian mass media. In *Proceedings of the 12th IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'20)*. IEEE Press, The Hague, Netherlands, 627–634. https://doi.org/10.1109/ASONAM49781.2020.9381344

[42] Mike Thelwall and Kevan Buckley. 2013. Topic-based sentiment analysis for the social web: The role of mood and issue-related words. *J. Amer. Soc. Info. Sci. Technol.* 64, 8 (2013), 1608–1617. https://doi.org/10.1002/asi.22872

[43] Teun van Dongen. 2021. Assessing the Threat of Covid 19-related Extremism in the West. Retrieved from https://icct.nl/publication/assessing-the-threat-of-covid-19-related-extremism-in-the-west-2/.

[44] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher Pal. 2016. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier Science & Technology, San Francisco, CA.

[45] Tomer Wullach, Amir Adler, and Einat Minkov. 2021. Towards hate speech detection at large via deep generative modeling. *IEEE Internet Comput.* 25, 2 (2021), 48–57. https://doi.org/10.1109/MIC.2020.3033161

[46] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., Vancouver, Canada, 17283–17297. Retrieved from https://proceedings.neurips.cc/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf.

[47] Jian Zhu, Zuoyu Tian, and Sandra Kübler. 2019. UM-IU@LING at SemEval-2019 task 6: Identifying offensive tweets using BERT and SVMs. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 788–795. https://doi.org/10.18653/v1/S19-2138