

DS_233: Model card

Data Science

Exported on 04/30/2024

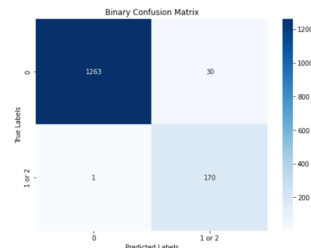
Table of Contents

1	Model Card for Safeguarding	3
2	Model at a glance.....	4
2.1	Intended Use	4
2.2	Metrics.....	5
2.3	Inputs and Outputs	5
3	Data.....	6
3.1	Training Data	6
3.2	Evaluation Data	7
4	Methodology	8
4.1	Training Methods.....	9
4.2	Evaluation	9
4.3	Quantitative Analyses	10
4.4	Ethical Considerations.....	12
4.5	Caveats and Recommendations.....	12
5	Model Deployment.....	14

1 Model Card for Safeguarding

A.K.A Suicidal and Self-harm Ideation Content Detector

2 Model at a glance

Description:	The Suicidal and Self-harm Ideation Content Detector (SSICD) is a model trained to identify content related to suicidal thoughts and self-harm in user reviews on the NHS.uk website. It was developed to enhance digital safety and provide timely support to vulnerable individuals accessing the service. The SSICD analyses text to flag content of self-harm or suicidal ideation for further human moderation.				
Model type:	Natural Language Processing (NLP), specifically using Transformer-based models (BERT) for text classification.				
Developed By:	NHS England's Data Science Skilled Team (@Liliana Valles Carrera and @Alice Tapper)	Confusion Matrix:		Recall:	0.99
Launch Date	📅 13 Jul 2023				Precision: 0.85
Version	1.0				F1 score: 0.92 Accuracy: 0.98

- **Development Background:** This model is one of a collection of models that have been designed by NHS England's Data Science Skilled Team (DSST) for the [nhs.uk](https://www.nhs.uk)¹ ratings and reviews (RAR) service, to replace human moderation of incoming reviews. The models have been designed to implement [RAR policy](#)², which defines a list of rules which reviews must pass in order to be published on the website. Developed in-house as part of a solution to identify safeguarding (suicidal ideation and self-harm) content in user reviews. Aimed to avoid the publishing of reviews with that content, and enabling expert moderators to further review and signpost the appropriate support to vulnerable individuals.

2.1 Intended Use

- **Scope:** Automatic identification of suicidal and self-harm ideation in written reviews on Rates and Reviews Pages on the NHS.uk domain. Classification of user reviews into 'no safeguarding', 'safeguarding low-risk', and 'safeguarding high-risk'.
- **Intended Users:** NHS.UK Rates and Reviews moderators.

¹ <http://nhs.uk/>

² <https://www.nhs.uk/our-policies/comments-policy/>

- **Use Cases Out of Scope:** Identification of other safeguarding issues like threats or violence not related to self-harm or suicide.

2.2 Metrics

- **Chosen Metrics:** Recall, F1 and accuracy
- **Rationale:** The model seek a high recall of the model. This metric measures the proportion of true positives detected by the model. A high recall ensures safeguarding instances are not missed, even at the price of having many false positives.

2.3 Inputs and Outputs

- **Inputs:** Short text (strings)

The endpoint in which the model has been deployed requires its input to be on the following format

Input Data
<code>{"data":submission_words}</code>

- **Outputs:**

The model outputs two key pieces of information:

- **Classification:** This is an integer value representing the category of safeguarding concern (0 for no safeguarding, 1 for low-risk, 2 for high-risk).
- **Probability:** This provides a confidence score for the classification, indicating how certain the model is of its prediction.

In the moderation process, the classifications of low-risk and high-risk are grouped. Consequently, the model effectively operates as a binary classifier, distinguishing between reviews that require safeguarding attention (either low or high risk) and those that do not.

The probability scores, while informative, are not utilized in the moderation process. They offer insight into the model's confidence but do not influence the binary classification outcome.

3 Data

- **Data Overview:** The model uses text data primarily sourced from user reviews about healthcare providers submitted on the [NHS.UK website](https://www.nhs.uk/contact-us/find-out-how-leave-review-of-nhs-service/)³. These reviews, stored in Dynamics CRM for up to two years, were subject to a rigorous two-level moderation process involving both external and internal teams. Due to historical inconsistencies in moderation, the internal team has curated the data to ensure accuracy and consistency. The dataset includes a mix of reviews: those that have breached moderation rules and those published without issues. It provides a comprehensive range of user feedback, crucial for training the model and evaluating its performance.

The data used to fine-tune the model comes from a mix of real historic comments on NHS UK rates and reviews page, *copycat* comments written by the expert moderators to simulate safeguarding high-risk comments, published reviews that do not include safeguarding comments and *augmented* data, that is, comments generated using NLP techniques based on real comments.

- **Data Pre-processing and cleaning:** There was no preprocessing to the data, a pretrained encoding model "uncased" was used, which means that it does not distinguish between *english* and *English*. The encoding requires an input form limited to 512 tokens: ([CLS] Sentence A [SEP] Sentence B [SEP]) [SEP] and [CLS] are special tokens that the model needs to function properly. The option `encode_plus` provided by the transformers library was used; it preprocesses the data as the classifier will need it. It tokenizes the review, adds the special characters for the model to identify spaces and padding (the model requires a specific length), maps the token to their IDs, truncates sentences above 512 and lastly, and it creates attention masks to differentiate real tokens from the empty spaces flagged with PAD.
 - **Encoding Model:** [bert base uncased](https://huggingface.co/bert-base-uncased)⁴
- **Split:** To evaluate the model's effectiveness, we split our data into:
 - Training set (~70%): Used for learning patterns and relationships
 - Validation set (~30%): Helps to understand its performance. No augmented data was used in the validation of the model

3.1 Training Data

- **Data Description:** Published and rejected reviews because of safeguarding content. The training data consisted of a subset of the overall data, specifically curated to include a balanced representation of reviews with and without safeguarding content and for the three risk levels. The data (text) was not preprocessed but was directly encoded using BERT encoding. This represents about 70% of all the historical data available at the time of training.

³ <https://www.nhs.uk/contact-us/find-out-how-leave-review-of-nhs-service/>

⁴ <https://huggingface.co/bert-base-uncased>

3.2 Evaluation Data

- **Datasets Used:** Published and rejected reviews because of safeguarding content, *copycat* comments (the in-house moderation team doctored some of the real historical reviews that were low risk, adding or changing phrases to make them high risk) and published reviews.
- **Why these datasets:** To cover a range of potential safeguarding scenarios, including low and high-risk reviews. There was scarcity of data with safeguarding content and augmenting and writing *look alike* data was useful to improve the performance and the confidence in the model.
- **Representativeness:** This represents about 30% of all the historical data available at the time of training, the other 70% was used for training.

Label	Origin	Training	Validation
No safeguarding	real	1,100	1293
Safeguarding	copycat	336	63
	generated	1,303	
	real	549	108
	Total	3,288	1464

4 Methodology

- **Model Type:** Natural Language Processing (NLP), specifically using Transformer-based models (BERT) for text classification.
 - **Algorithm Details:**
 - **Algorithm Used:** BERT (Bidirectional Encoder Representations from Transformers), a pre-trained language model designed for natural language understanding.
 - **Models used:** BERT serves as both the feature extractor and the classification model: it transforms text into contextualized embeddings (sentence embedding vectors) which are then used by the same BERT architecture, pre-trained for classification tasks. The model weights were then fine-tuned using our labelled data, consisting of both safeguarding and non-safeguarding texts. [bert base uncased](https://huggingface.co/bert-base-uncased)⁵ is a deep neural network with 12 layers, trained by Google on Wikipedia and Book Corpus (a dataset with more than 10,000 books of different genres). It produces a word representation that is dynamically informed by the words around them. It is a transfer learning model capable of understanding complex language, patterns and sentiments while coping with unseen words that are not part of its 30,000 word vocabulary.
 - **Fine-Tuning Approach:** The model is fine-tuned on a dataset specifically labelled for suicidal and self-harm ideation content, allowing it to learn the special characteristics of this type of text.
 - **Classification Layer:** A classification layer is added on top of the pre-trained BERT model to categorize text into relevant classes (e.g., 'safeguarding', 'no safeguarding').
 - **Feature Engineering:** Text embeddings (of the review left by the user) were used as the only feature as they provide a rich representation of the semantic content and contextual details of the review.
 - **Training Process:** A feature-based approach with a pre-trained BERT model was used, that is, text is converted into numerical vector representations (BERT as encoding). These vectors are then used as inputs for a classification task. The core of the training involves fine-tuning the pre-trained BERT model on a dataset that includes both safeguarding and non-safeguarding texts. This process adjusts the model's weights (of the neuron network) to effectively differentiate between texts that indicate potential self-harm or suicidal ideation and those that do not. We modified a pre-trained BERT model using the huggingface PyTorch implementation interface designed for Classification (BertForSequenceClassification). The full list of classes or interfaces for fine-tuning that BERT provide is documented here: [BERT](https://huggingface.co/docs/transformers/en/main_classes/tokenizer)⁶. The way it works is by adding a layer on top for the classification task.
 - **Hyperparameter Tuning:** The parameters to tune were Batch size, learning rate and number of epochs. The values tested for these parameters were the ones suggested in the original [BERT paper](https://arxiv.org/pdf/1810.04805.pdf)⁷ (BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding by Jacob Devlin), and the final combination was chosen based on the results obtained on a validation set (independent

⁵ <https://huggingface.co/bert-base-uncased>

⁶ <http://se%20more%20about%20the%20tokenizer%20used%20here%20tokenizer/>

⁷ <https://arxiv.org/pdf/1810.04805.pdf>

from the train and test datasets) and the processing limitations of Azure Machine Learning (it didn't cope with batch sizes above 8). The model ran for 4 epochs with a learning rate of $3e-5$ and a batch size of 4. It was selected as it was the one with the highest accuracy on the validation holdout that minimises the false positives and false negatives.

- **Alternate Methods:** Alongside the primary BERT model, we evaluated alternative methods for the SSICD, including encoding models like E5 and BGE, and classifiers such as XGBoost, Logistic Regression, and SVM. These methods were tested for their ability to process text (the time it took to encode the text) and classify content accurately. Although they presented simpler approaches compared to BERT, the deep learning capabilities of BERT ultimately showed superior performance.

4.1 Training Methods

- **Literature Review:** A literature review was conducted focused on existing methods for identifying suicidal ideation in short texts like social media. A systematic review titled [Suicidal Ideation Detection on Social Media: A Review of Machine Learning Methods](#)⁸ was particularly useful as it compared the sources, annotation, feature selection, and classification methods used in multiple research and provided their performance metrics. This informed the selection of using classification methods and the text content as the feature. BERT was not used by any of the researchers, but given that it was a recent development in natural language processing, we compared its performance over other classification algorithms and embedding models.
- **Preferred Approach:** BERT-based models for superior performance

4.2 Evaluation

- **Model Validation:**
The model was tested on a holdout (validation) dataset, which is a separate set of data not used during the training phase. It represents the various levels of risk of suicidal and self-harm ideation (no risk, low-risk and high-risk).
Following advice from Clinicians, we prioritized minimizing false negatives due to the sensitive nature of the content. A false negative, where a genuine case of suicidal or self-harm ideation is not identified, could have serious implications for user safety so, while maintaining a balance, we leaned towards a higher tolerance for false positives. This strategy ensures that the model is more likely to flag content that requires further human review.

The validation process also involved evaluating the key performance metrics precision and F1 score.

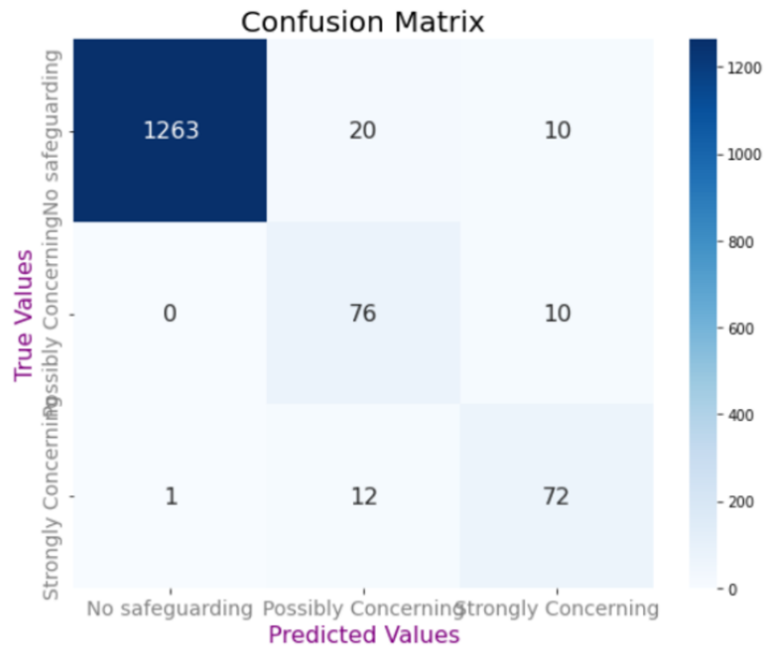
- **Evaluation Focus:** The cost of a false positive is that it causes annoyance to the user and prevents them from publishing the review immediately as it will be sent to a human moderation for review. The cost of a false negative allows the publication of a potentially sensitive review that can be upsetting or even triggering for other readers. Therefore, while keeping a high level of accuracy, there is a preference for a high recall.

⁸ <https://arxiv.org/abs/2201.10515>

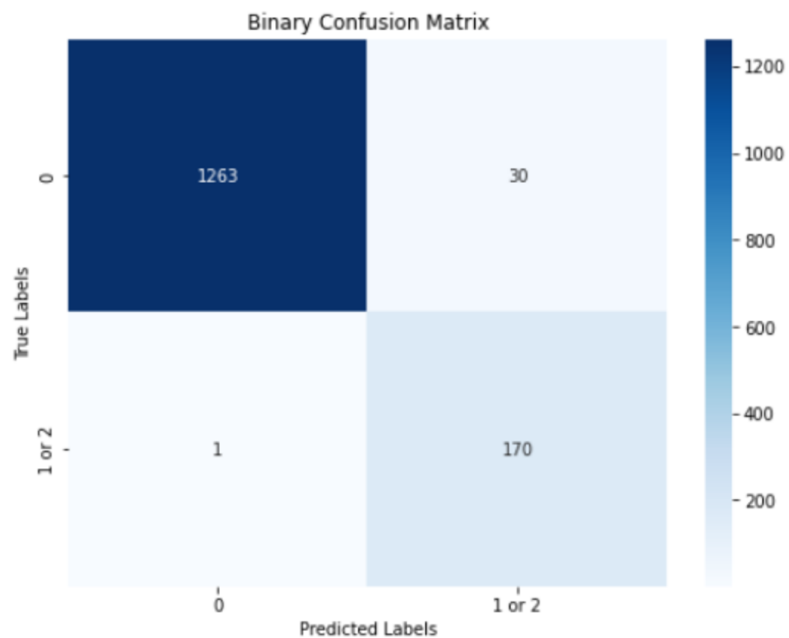
4.3 Quantitative Analyses

- **Performance Breakdown:** The model's performance was not specifically broken down by demographic or other characteristics due to the unavailability of such data in the review content. No additional characteristics were identified as relevant that warranted separate performance tracking.
- **Model Validation:**

The accuracy is 0.96
 The F1 score is 0.97
 The precision is 0.97
 The recall is 0.96



The accuracy of the binary classifier is 0.98
 The F1 score of the binary classifier is 0.92
 The precision of the binary classifier is 0.85
 The recall of the binary classifier is 0.99



- **Bias and fairness analysis:**

A bias analysis was conducted on the three models trained by the team for the automoderation tool: Safeguarding. The analysis aimed to evaluate the fairness of these models by examining the sentiment of misclassified comments, the length of the texts, and the number of spelling mistakes standardized by text length, to identify any potential bias in the models.

The models operate in contexts where demographic data is not accessible, relying solely on textual content to make classifications. As such, it is crucial to ensure that these models do not inadvertently favour or penalize certain characteristics that could lead to biased outcomes. For instance, longer texts might be inherently more complex or contain more nuanced sentiment, affecting their classification. Similarly, the frequency of spelling mistakes is a proxy for the literacy levels of the writer and we definitely don't want to favour, for example, positive reviews over negative ones. It was observed that longer texts are more prone to misclassification and more negative or neutral comments tend to be wrongly flagged as containing safeguarding than those slightly more positive. This is not surprising given the negative sentiment of the safeguarding issues.

See [DS_233: Bias analysis](#)⁹ results.

4.4 Ethical Considerations

- **Sensitive Data Use:** The handling of reviews involves dealing with potentially sensitive information, which includes suicidal ideation or self-harm, either for training, validation, analysis or moderation. This data is stored with high-security measures either on Dynamics or the Azure datastore only accessible given the right privileges and permissions. This is to ensure confidentiality and privacy.
- **Implications for human safety:**
 - There is no direct impact on the well-being and safety of individuals, this is not a medical device. However, there's a risk of false negatives, which could result in failing to identify a user in need of urgent help, a missed opportunity for timely intervention. However, the model is not used as a standalone tool but as part of the Automoderation tool, where the likelihood of receiving a review with safeguarding content is only 1 every 1000. All the reviews flagged by the model as either low risk or high risk are further reviewed by an expert moderator who, after confirming the classification, will reach out to the review author or refer to qualified clinicians. If the safeguarding content is not identified by the model, the review will be published, but the provider receiving the review will be alerted and instructed to flag back to the rates and reviews team if they do identify such content.
 - The main risk is the potential for false positives and negatives, which could lead to unnecessary distress for users or missed opportunities for intervention. Regular reviews and updates of the model are planned to mitigate these risks.

4.5 Caveats and Recommendations

- **Caveats and limitations:** The model relies only on the text included in the reviews and might miss nuances in the English language including new cultural references, new idioms, complex figurative language, extreme sarcasm and irony and ambiguous sentences.

⁹ https://nhsd-confluence.digital.nhs.uk/display/DAT/DS_233%3A+Bias+analysis

- **Additional Notes:** Users of the Rates and Reviews platform are informed of the use of automated tools to assess their compliance with the moderation policy.
- **Future Directions:** Refining the model to include a wider range of vocabulary, it was observed that the model fails to understand some modern/colloquial words; this is not a big concern as it is not the typical language used to write reviews. A more relevant improvement would be to tackle the bias on the length of the reviews where longer reviews tend to be misclassified.

5 Model Deployment

- **Versioning:** 1.0 registered as 'safeguarding_bert_bert3cat:6'
- **Links:** The model and its associated notebooks are stored in a secure Azure repository, accessible only to authorized members of the NHS England's Data Science Team.
- **Inference Process:** The inference process for the Safeguarding model involves receiving and processing user reviews in real-time through a Flask app, which queries the Azure-hosted model endpoint. The model promptly classifies the reviews for potential safeguarding concerns, with the results immediately directed to human moderators for further analysis.
- **Deployment Strategy:**
 The model, developed on Azure Machine Learning, leverages Azure's managed services. The models are deployed using the Standard_DS3_v2 SKU, a general-purpose endpoint type that provides a balance of CPU, memory, and I/O, suitable for various workloads. This SKU features 4 cores and 14 GiB of RAM. The deployment allows for both manual and automatic scaling, with autoscaling as a built-in feature. The current setup, with an instance count of 1, is optimized for the anticipated workload but can be adjusted based on demand. Monitoring performance and response times dashboards were set up to determine the need for scaling, either by increasing the number of instances or upgrading to a more powerful SKU.
 The compute resources are managed by Azure, simplifying scaling and management processes. The endpoint is secured with key-based authentication to ensure only authorized access.
 The Safeguarding model's endpoint, specifically, is available at <https://safeguarding-redeployed.uksouth.inference.ml.azure.com/score>. It integrates with the rates and reviews infrastructure of NHS.UK. The model is queried through a Flask app, which calls the model's endpoint, enabling immediate moderation of user-submitted reviews on the rates and reviews webpage. The responses are then sent to human moderators for further analysis.
- **Model Performance in Deployment:** Ongoing monitoring and evaluation are in place to ensure the model's performance remains consistent and reliable, with periodic updates as needed. An analysis was performed on the data collected between mid-July when the model was first launched and the 6th of December 2023. The results are recorded here: [Safeguarding analysis](#)¹⁰
- **Timing Analysis:** Some general timing analysis was performed to see how quickly the Flask app can query, receive and handle the results from each endpoint, documented [here](#)¹¹. This can be used as an indirect measure of the endpoint latency for practical purposes.

¹⁰ <https://nhsd-confluence.digital.nhs.uk/display/DAT/Safeguarding+analysis>

¹¹ https://nhsd-confluence.digital.nhs.uk/display/DAT/DS_233%3A+Performance+Testing