

# **DS\_233: Modelcard Complaints**

Data Science

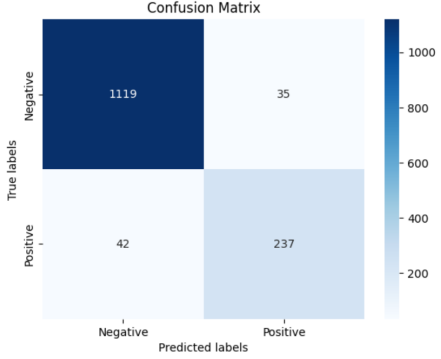
Exported on 04/30/2024

# Table of Contents

1	Model Card for Complaints .....	3
2	Model at a glance.....	4
2.1	Intended Use .....	4
2.2	Inputs and Outputs .....	5
3	Data.....	6
3.1	Training Data .....	6
3.2	Evaluation Data .....	6
4	Methodology .....	7
4.1	Training Methods.....	7
4.2	Evaluation .....	7
4.3	Quantitative Analyses .....	8
4.4	Ethical Considerations.....	9
4.5	Caveats and Recommendations.....	10
5	Model Deployment.....	11

# 1 Model Card for Complaints

## 2 Model at a glance

<b>Description:</b>	The Complaints model was trained to identify written reviews that violate the NHS complaints policy by detecting content that should be escalated to formal complaints rather than addressed through the review process on the NHS.UK ratings and reviews service. It aims to serve as the first line of automated moderation for the service. It analyses text to distinguish between a given review ought to be raised as formal compliant or not, according to the NHS.UK moderation policy.				
<b>Model type:</b>	Natural Language Processing (NLP), a classifier model using a supervised learning approach				
<b>Developed By:</b>	NHS England's Data Science Skilled Team (@Daniel Goldwater)	<b>Confusion Matrix:</b>	 <p>Confusion Matrix</p> <p>True labels \ Predicted labels</p> <p>Negative      Positive</p> <p>Negative      Positive</p>	<b>F1 score:</b>	0.86
<b>Launch Date</b>	📅 22 Dec 2023			<b>Precision:</b>	0.87
<b>Version</b>	1.0			<b>Recall:</b>	0.85

- **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)<sup>1</sup> ratings and reviews (RAR) service, to replace human moderation of incoming reviews. The models have been designed to implement [RAR policy](#)<sup>2</sup>, 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 complaints

### 2.1 Intended Use

- **Scope:** Automatic identification of Complaints in written reviews on Rates and Reviews Pages on the [NHS.uk](https://www.nhs.uk)<sup>3</sup> domain.

It is considered a complaint within the RAR service review if the content reflects any of the following:

1. Intent to make a complaint.

<sup>1</sup> <http://nhs.uk/>

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

<sup>3</sup> <http://NHS.uk>

2. They have made a complaint.
  3. Went through the complaint process and are still unhappy
  4. The review contains a serious accusation, such as:
    - Threats of violence
    - Medical negligence
    - Defamation
    - Libellous
    - Abuse
    - Rudeness
    - Misdiagnosis
    - Any allegation of crime
    - Crime taking place
- **Intended Users:** Patients leaving a review through the nhs.uk RAR service
  - **Use Cases Out of Scope:** Broad applications beyond the specific context of NHS ratings and reviews.

## 2.2 Inputs and Outputs

- **Inputs:** Short text (strings)
- **Outputs:** Classification results ("0" or "1") indicating whether a review violates the NHS complaints policy.

## 3 Data

- **Data Overview:** The model uses text data primarily sourced from user reviews about healthcare providers submitted on the [NHS.UK website](#)<sup>4</sup>. 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. Trained on real-world data gathered before January 2023, including reviews flagged as complaints and augmented datasets to improve model robustness.
- **Data Pre-processing and cleaning:** The text was cleaned in a straightforward way - stripping line endings, extra punctuation, and so forth. The positive label (complaints) data was reviewed by the second-line moderators.
  - **Encoding Model:** We used a version of [MNet](#)<sup>5</sup>; a pre-trained model specifically designed for sentence embedding tasks like **semantic search and clustering**. Specifically, we used [all-mpnet-base-v2](#)<sup>6</sup>.
- **Split:** The data set was split into train, test and validation: The validation (hold out dataset) was 98 rows, and the remaining 700 rows of the dataset were divided into training (75%) and test set (25%).

### 3.1 Training Data

- **Data Description:** This model was trained on real-world data gathered before January 2023. We also used data augmentation techniques to expand the training range. This included:
  - Sentence shuffling
  - Synonym substitution

### 3.2 Evaluation Data

- **Datasets Used:** Published and rejected reviews entirely gathered after the training and deployment of this model.
- **Why These Datasets:** It is a combination of original and AI-augmented data to balance the dataset to address the imbalance in the original dataset and ensure a robust model.
- **Representativeness:** All historical data available is expected to cover a range of scenarios for where a complaint according to our definition is included.

---

<sup>4</sup> <https://www.nhs.uk/contact-us/find-out-how-leave-review-of-nhs-service/>

<sup>5</sup> [https://huggingface.co/docs/transformers/model\\_doc/mpnet](https://huggingface.co/docs/transformers/model_doc/mpnet)

<sup>6</sup> <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

## 4 Methodology

- **Model Type:** Supervised learning classification model
  - **Models used:** We used [all-mpnet-base-v2](https://huggingface.co/sentence-transformers/all-mpnet-base-v2)<sup>7</sup> to generate the text embeddings. We then used an [SKLearn SVM](https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html)<sup>8</sup> model to classify those embeddings according to label.
  - **Algorithm Details:**
- **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 two-stage approach was taken, starting with the MPNet model, a transformer-based neural network was utilized to encode the input data and convert the review text into a high-dimensional vector representation. This representation was then used as the input for the SVM model. During training, multiple embedding models were
- **Hyperparameter Tuning:** Using hyperopt the C (Regularization parameter), gamma and epsilon, were optimized for an asymmetric goal prioritizing the reduction of false positives. We also optimised the amounts of each type of augmented data to use.
- **Alternate Methods:** During training, multiple embedding models were tested along with the following classification algorithms, with SVM providing a superior performance: KNeighborsClassifier, MLPClassifier, BaggingClassifier, RandomForestClassifier, AdaBoostClassifier, ExtraTreesClassifier, GradientBoostingClassifier, NearestCentroid, GaussianNB, BernoulliNB, LogisticRegression, RidgeClassifier, SGDClassifier and XGBoost.

### 4.1 Training Methods

- **Literature Review:** No literature review was carried out as the definition of complaints used by the service is a bespoke one and no previous attempts were made to automatically flag those reviews.
- **Preferred Approach:** The selected approach provided the expected performance and it is consistent with other modeling approaches used for the automoderation tool.
- **Encoding and Fine-Tuning Details:** A pipeline was developed to allow testing the combination of 6 embedding models with multiple classifiers, their parameters optimised and the best combination was selected.

### 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 two classes: complaint, no complaint. The

<sup>7</sup> <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

<sup>8</sup> <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>

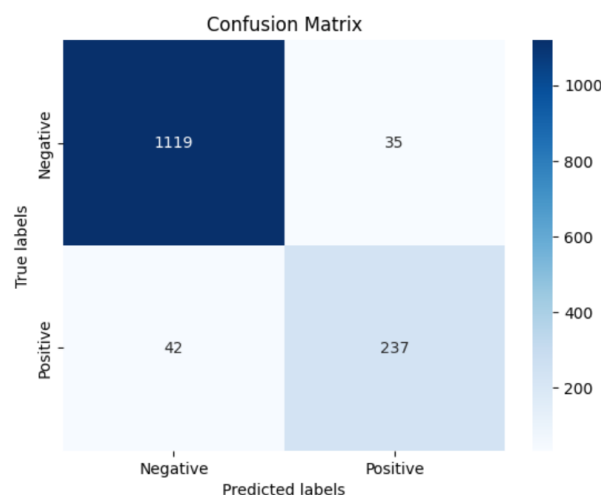
performance was also compared with that of the first line of human moderators and the model performed better.

- **Evaluation Focus:** We focused on reducing the numbers of false positives, that is, to not flagging reviews as complaints when they are not because the complaints process generates a cost to the NHS. However, we also wanted to avoid missing real complaints. Given the emphasis on reducing false positives without significantly compromising the ability to detect true complaints, a somehow balanced approach to precision and recall is desirable. However, since the model prioritizes minimising false positives (to avoid unnecessary costs), a high precision is more critical than recall.

## 4.3 Quantitative Analyses

- **Model Validation:** The model was tested on a holdout (validation) dataset, which is a separate set of data not used during the training phase. Though we did of course have a validation dataset at the time of the model's creation, the results presented on this page come from data collected entirely after the model's creation and deployment.
- Recall: 0.85
- Precision: 0.87
- F1 Score: 0.86

Here is a confusion matrix showing the model's performance against data gathered after the model was trained. A key and explanation follow below.



\*These labels translate as

In figure	From Model	Meaning
'Negative'	0	No complaint detected
'Positive'	1	Complaint detected



When interpreting these results, bear in mind that the data is highly imbalanced. The ratio of imbalance here does not perfectly resemble the imbalance in the real-world incoming data, but does lean in the same direction.

- **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.
- **Bias and fairness Analysis:** A bias analysis was conducted on the three models trained by the team for the automoderation tool including Complaints. 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.
- In training models for this project, and specifically this model, we use only the relevant fields from submitted reviews. Specifically, our models are trained on `Comment Title` and `Comment Text`. All other context (name, D.O.B., demographic identifiers, organisation they are reviewing for, etc) are excluded. This helps prevent against models forming biases based on other variables by proxy, and works towards ensuring that our models are achieving the results to collect because of the content of the reviews they are categorising, and nothing else.
- 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.

The average length of false positives (texts incorrectly classified as complaints) is greater than that of true negatives (texts correctly classified as non-complaints), that is, the model is more prone to incorrectly flag longer texts as complaints. The model appears to be better at identifying texts with fewer spelling mistakes as complaints. This could be due to the model associating a lower spelling mistake rate with more formal or carefully written complaints. Reviews (complaints) that are misclassified as non-complaints (FN) have a higher spelling mistake rate, which could be a contributing factor to their misclassification. Negative comments tend slightly to be classified as complaints, a trait expected of this type of comment. The size effect is medium. See [DS\\_233: Bias analysis](#)<sup>9</sup> results.

## 4.4 Ethical Considerations

- **Sensitive Data Use:** 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.
- **Risks including implications for human safety:** the primary risks concerning human safety and potential harm are minimal given the model's scope and application. However, there can be indirect implications for individuals whose complaints may not be accurately identified by the model. If genuine complaints are missed (false negatives), a response to serious concerns raised in the reviews can still be picked up by the providers receiving the complaint.

---

<sup>9</sup> [https://nhsd-confluence.digital.nhs.uk/display/DAT/DS\\_233%3A+Bias+analysis](https://nhsd-confluence.digital.nhs.uk/display/DAT/DS_233%3A+Bias+analysis)

## 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.
- **Future Directions:** It is suggested to keep a training and evaluation process to detect any shift in the language.
- **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.

## 5 Model Deployment

- **Versioning:** 1.0 deployed [here](#)<sup>10</sup> and registered [here](#)<sup>11</sup>.
- **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 Complaints 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 complaints.
- **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.

- **Model Performance in Deployment:** \*\*To be updated once data is collected\*\*
- **Timing Analysis:** Inference times measured for different approaches, with a focus on handling longer texts.

---

<sup>10</sup><https://ml.azure.com/endpoints/realtime/complaints-multimodel/detail?wsid=/subscriptions/07748954-52d6-46ce-95e6-2701bfc715b4/resourcegroups/nhsuk-automoderation-rg-dev-uks/providers/Microsoft.MachineLearningServices/workspaces/nhsuk-automodapi-ml-dev-uks&tid=d9fec63b-47ed-4fa5-a265-1d3bb934a78b>

<sup>11</sup><https://ml.azure.com/model/minimal-multimodel:8/artifacts?wsid=/subscriptions/07748954-52d6-46ce-95e6-2701bfc715b4/resourcegroups/nhsuk-automoderation-rg-dev-uks/providers/Microsoft.MachineLearningServices/workspaces/nhsuk-automodapi-ml-dev-uks&tid=d9fec63b-47ed-4fa5-a265-1d3bb934a78b>