**Supplementary Material**

**Model and parameter search**

Table S1 lists the included models and their respective hyperparameters. Table S2 lists the preprocessing hyperparameters. All hyperparameters were searched using a grid search. Hyperparameter values in bold were used for fitting the final model. Note that Naïve Bayes provides a strong baseline model (Zhang, 2004) and that optimization over only a few key learning algorithms is often preferable as it limits the search space and allows for easier validation, maintenance, and explainability. Given a limited compute budget only a subset of valid hyperparameters were examined and slightly better performance (classification) could potentially have been obtained by conducting a wider grid search. Note, however, that the current grid search resulted in more that 9000 fits. Specifically, for the XGBoost, 324 (3×3×3×3×4) different models were fit for each of the preprocessing hyperparameters resulting in 1944 (324×3×2) fits, which were then cross validated five times (resulting in a total of 9720 fits). For the Naïve Bayes, a total of 30 (1×3×2×5) fits were made as it has no hyperparameters.

**Table S1.** Model and their hyperparameters

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Hyperparameters** | **Potential Values** | **Highest obtained test AUC** |
| XGBoost | Learning rate | 0.1, 0.2, **0.3** | 0.8875 |
| Number of estimators | 50, 100, **200** |
| Maximum depth | **3**, 6, 12, 18 |
| Minimum child weight | 1, 2, **3** |
| Gamma | 0.0, **0.1**, 0.2 |
| Naïve Bayes | N/A | N/A | 0.8319 |

*Values in bold mark those achieving the highest AUC.* N/A: not applicable.

**Table S2.** Preprocessing hyperparameters

|  |  |
| --- | --- |
| **Hyperparameters** | **Potential Values** |
| Transformation | **TF-IDF** |
| N-gram range | (1, 2) |
|  | (1, 3) |
|  | **(1, 4)** |
| Minimum number of words | 2 |
|  | **5** |

*Values in bold mark those achieving the highest AUC.*

N-grams which appeared in more than 50% of the documents were removed as these are unlikely to be discriminative. Similarly, hapaxes (words which only appear once), were removed as the estimated effect of these variables is very uncertain. Before validating the model in the test set, we also experimented with over and under sampling as well as Synthetic Minority Oversampling Technique (SMOTE), but found no substantial effect of these steps for the XGBoost model using cross-validation. The Naive Bayes model, however, obtained a slightly higher performance (AUC=0.85) using SMOTE.

Model training and selection was performed in Python (version 3.7.9) using scikit-learn (Buitinck et al., 2013, version 0.23.2[)](https://www.zotero.org/google-docs/?zW1Iam), and imblearn (Lemaître et al., 2017, version 0.7.0). For XGBoost, the Python package xgboost was used (Chen & Guestrin, 2016, version 1.0.2).

**Model application and validation**



**Figure S1.** Histogram showing the distribution of the 500 random samples over time (grey) and the total number of misclassifications (red).

**References**

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). API design for machine learning software: Experiences from the scikit-learn project. *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108–122.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, *18*(17), 1–5.

Rohde, C., Jefsen, O. H., Nørremark, B., Danielsen, A. A., & Østergaard, S. D. (2020). Psychiatric symptoms related to the COVID-19 pandemic. *Acta Neuropsychiatrica*, *32*(5), 274–276. <https://doi.org/10.1017/neu.2020.24>

Zhang, H. (2004). The optimality of naive Bayes. *AA*, *1*(2), 3.