EIOS
EPIDEMIC INTELLIGENCE
FROM OPEN SOURCES

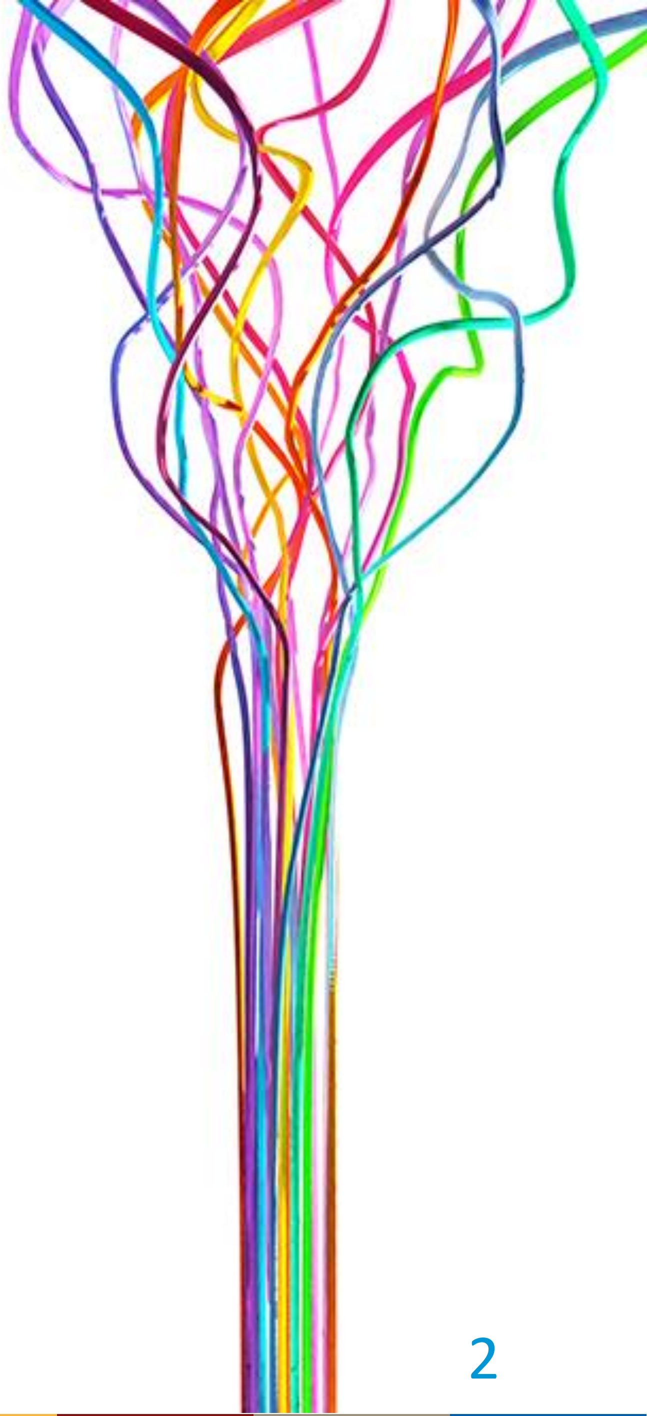
Using human expertise and text classification algorithms to identify the noise in EIOS

Scott Lee, Emilie Peron

EIOS Global Technical Meeting, 13 November 2019
Seoul, South Korea

Objectives

1. To **estimate** the proportion of noise in biological categories from the WHO and Global Health Security Initiative trees and the correlation coefficients
2. To **describe** the categories, the sources and the language associated with noise
3. To **estimate** the inter-expert agreement on noise classification by using the Cohen’s kappa score
4. To **train** a machine learning model to label the reports as noise or not noise

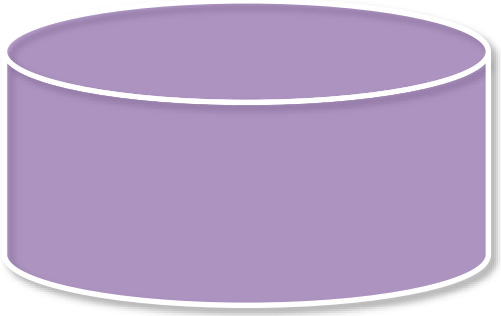


Information classification process in EIOS

MEDIA REPORTS SCREENED BY EIOS

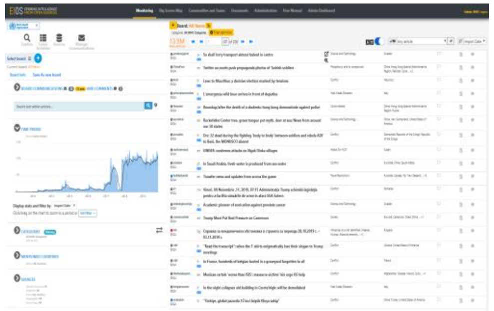
Categorisation
Deduplication
Entity recognition

EXCLUDED



TRUE NEGATIVE
FALSE NEGATIVE

IMPORTED

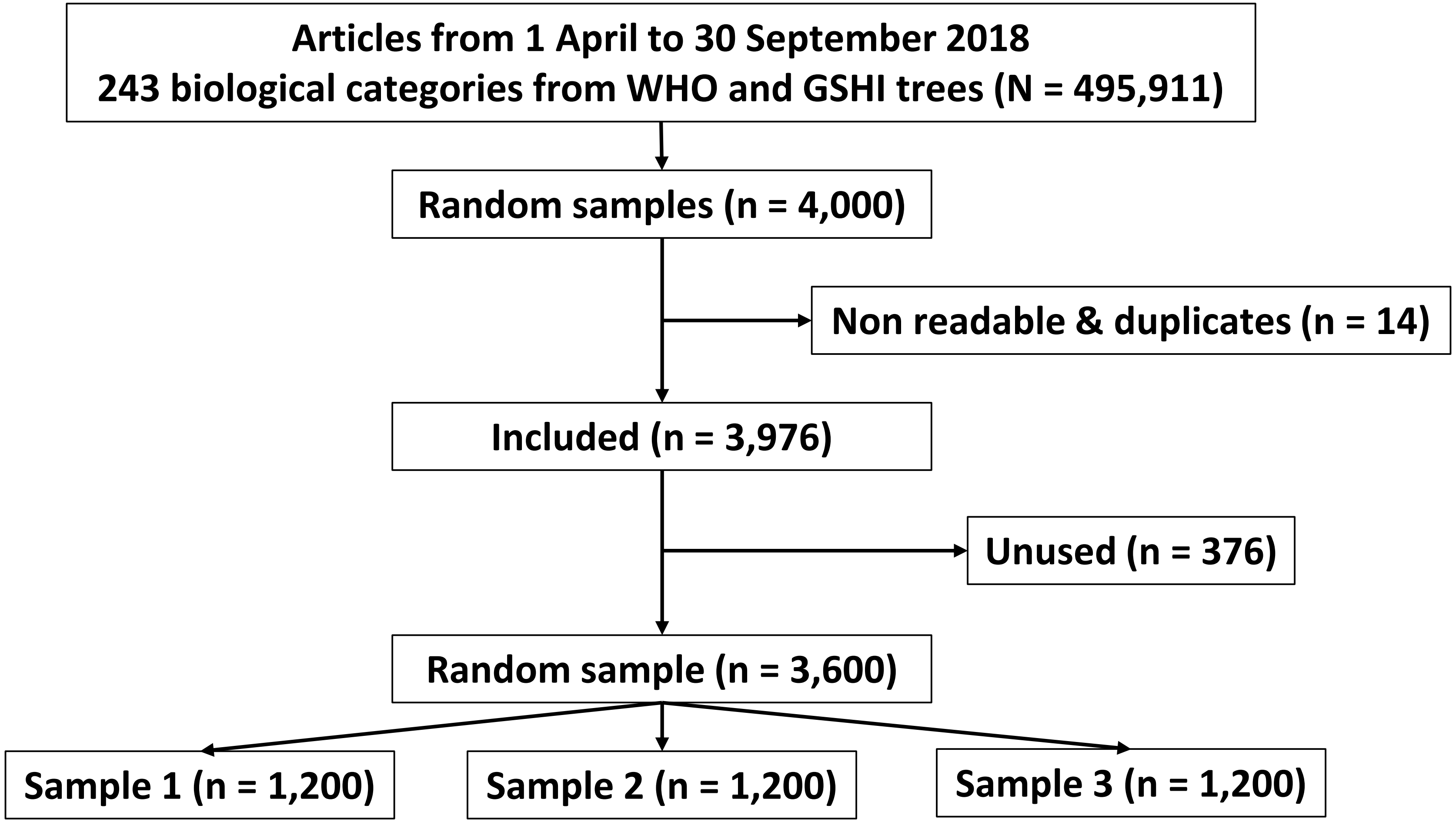


TRUE POSITIF
FALSE POSITIF
= NOISE

Relevant information definition

- Any public health-related information in an all-hazard and one-health approaches to:
  - **Detect** new a signal of interest that could have an impact on public health
  - **Add contextual and background information** that might change the risk assessment or lead to an action or contribute to the situational awareness
  - **Provide updated information** on a public health event that might change the risk assessment or lead to an action or contribute to the situational awareness
- Any information related to travel and trade ban

Random selection of the articles



Methods (1)

- Each sample sent to two paired experts for independent pre-classification (noise, not noise, I don't know)
- Experts from: ECDC, FAO, GOARN, GPHIN, OIE and US CDC
- Comparison of the pre-classification to extract the discrepant records
- Review of the discrepant records and final classification by two experts at WHO
- Univariable analysis: intermediate & final Cohen's Kappa score, overall percentage of noise by category, source and language

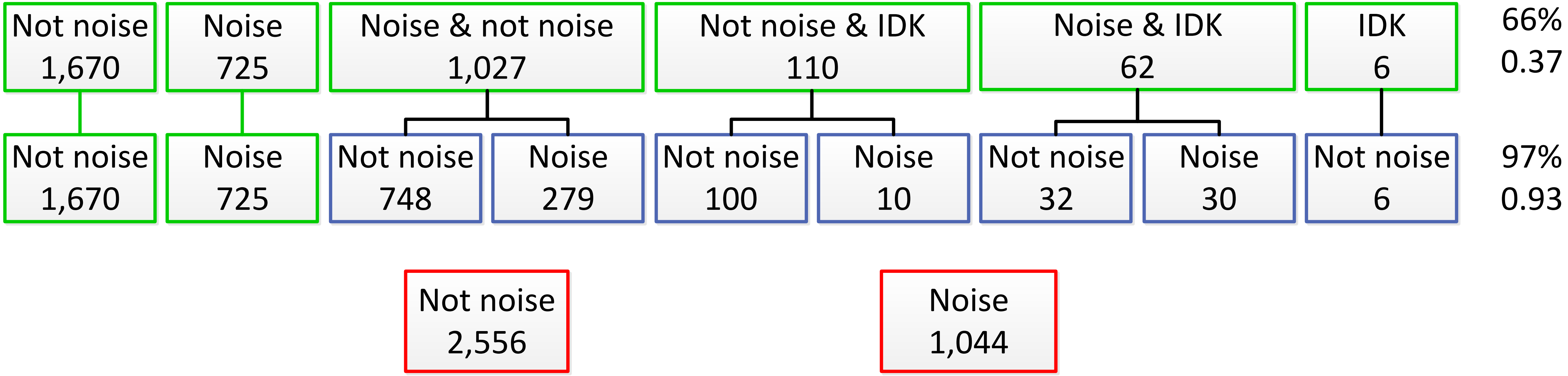
Methods (2)

- Tried a variety of models, including a random forest, a support vector machine, and BERT, a deep neural network from Google

- Monte Carlo cross validation, or randomly splitting the dataset into training and test sets a bunch of times (100 splits)

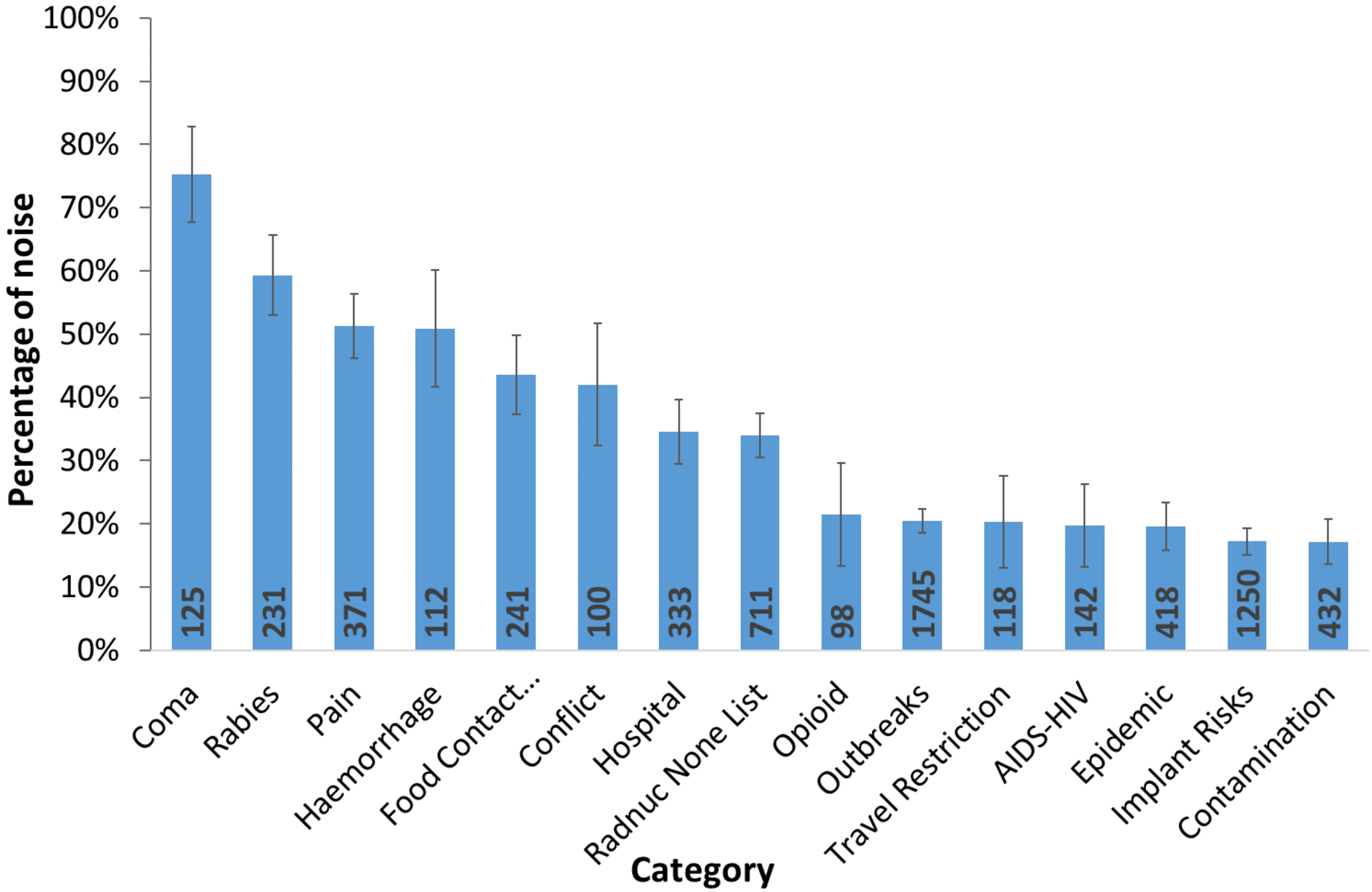
- Measures of binary classification accuracy, like sensitivity and F1 score; and measures of probabilistic accuracy, like Brier score loss

Noise proportion, Cohen's kappa score

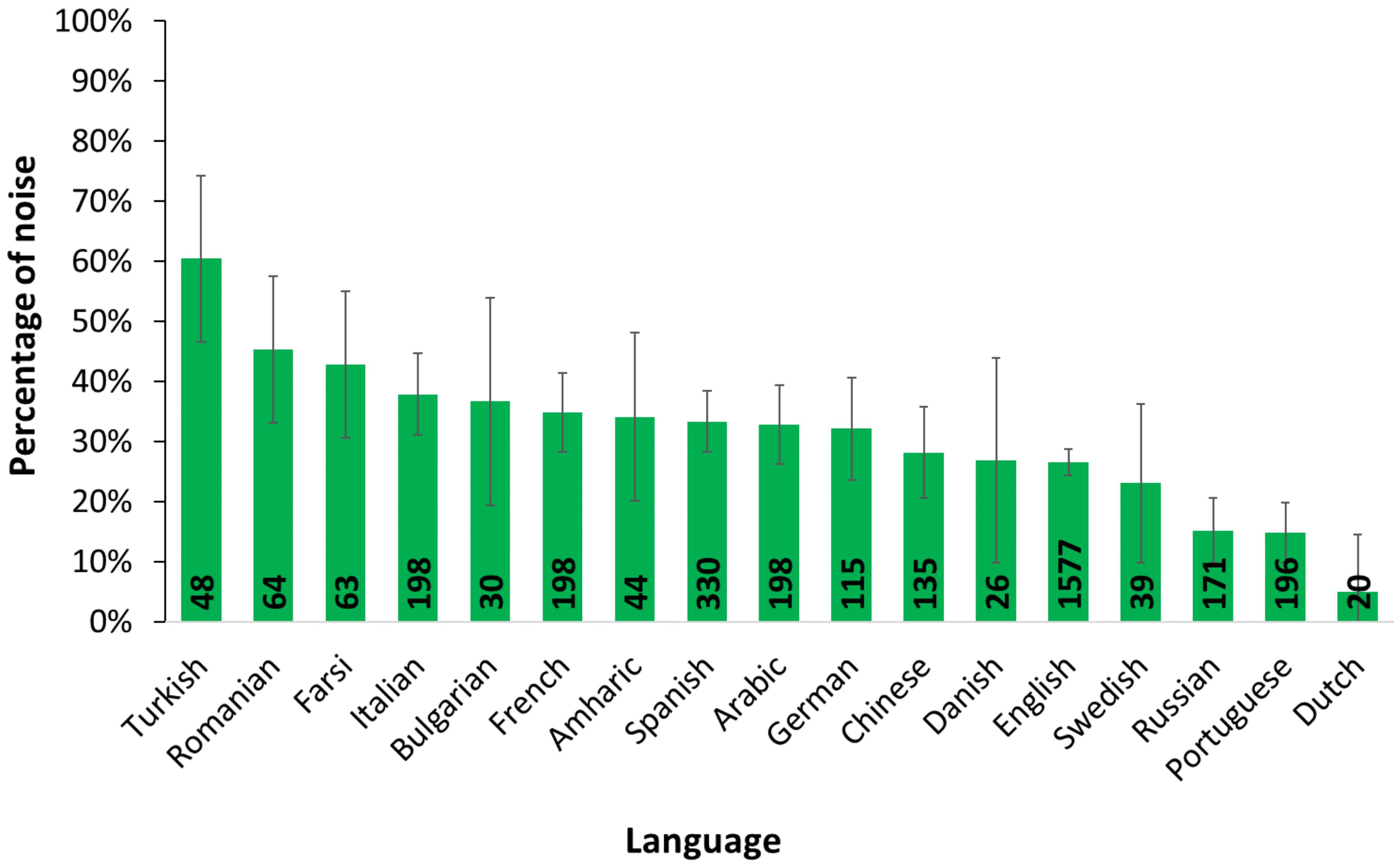


Proportion of noise = 29% (95%IC 27% to 31%)

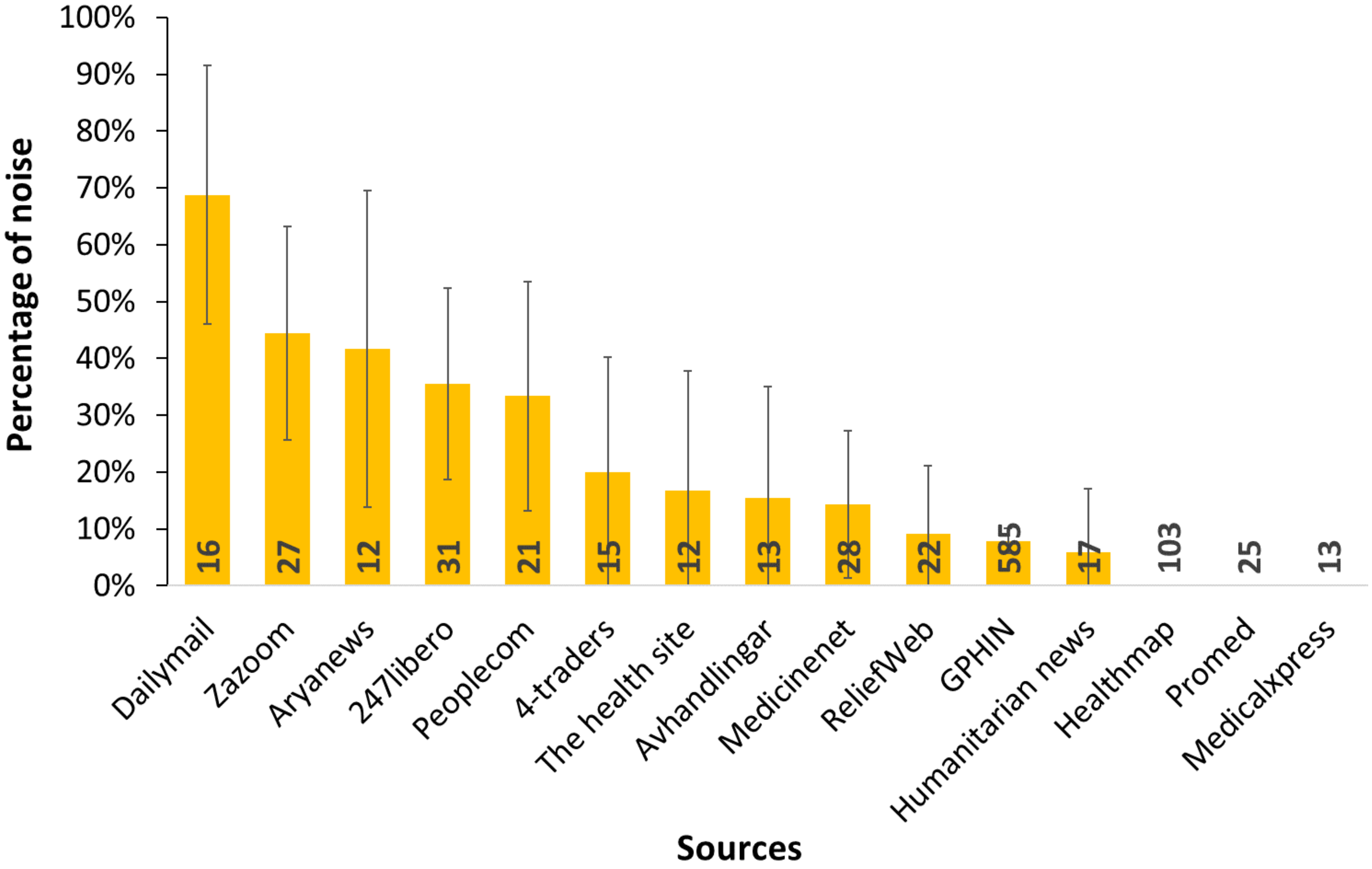
Noise proportion in the 15 most common categories



Noise proportion by languages with at least 20 articles



Noise proportion in the 15 most common sources



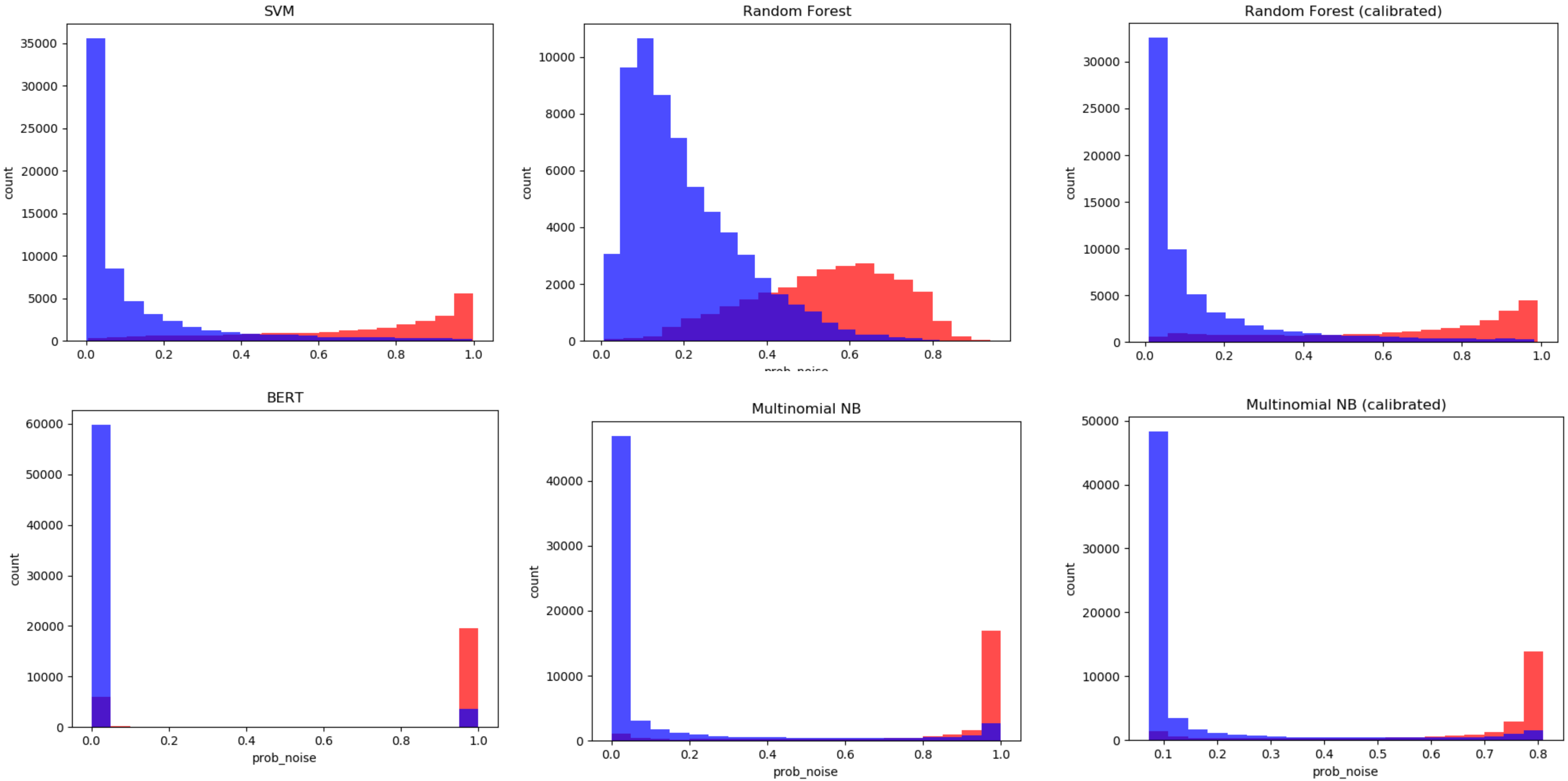
Text classification

Multinomial Naïve Bayes wins on sensitivity (86%), NPV (94%), and F1 (80%)
Platt-calibrated MNB wins when targeting prevalence (1% discordant)
SVM is best-calibrated overall (Brier = .083)

mod	tp	fp	tn	fn	sens	spec	ppv	npv	f1	pos	pred_pos	diff	rel	mcn.p	brier
BERT	198	38	601	63	0.7585	0.9403	0.8401	0.9054	0.7962	261	236	-25	-0.0954	0.1307	0.1100
rf	160	25	614	101	0.6134	0.9604	0.8639	0.8589	0.7169	261	185	-76	-0.2897	0.0000	0.1131
rf_cal	185	45	594	76	0.7075	0.9295	0.8044	0.8862	0.7525	261	230	-31	-0.1200	0.0360	0.0979
svm	198	42	597	63	0.7594	0.9341	0.8252	0.9049	0.7906	261	240	-21	-0.0793	0.1361	0.0833
mnb	224	72	567	37	0.8581	0.8875	0.7575	0.9388	0.8045	261	296	35	0.1334	0.0181	0.0953
mnb_cal	210	54	585	51	0.8055	0.9148	0.7947	0.9202	0.7998	261	265	4	0.0142	0.5267	0.0891

Mean performance for each model across the 100 random train-test splits.

Distributions of predicted probabilities

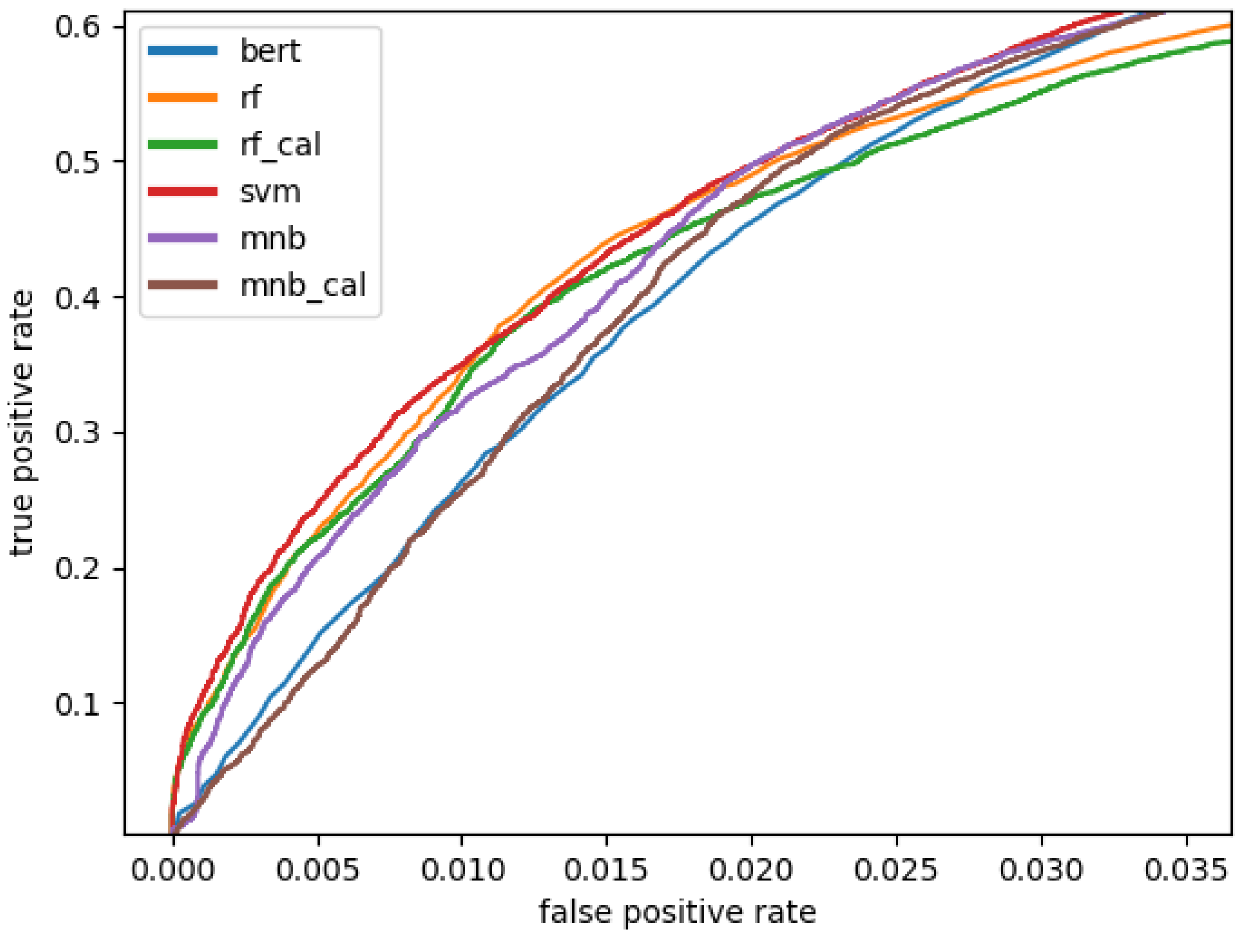


ROC curve analysis

When false positive rates (FPRs) are low, the **calibrated SVM** consistently performs best, detecting 10%, 35%, and 85% of the noise reports at different FPRs.

BERT was nearly the worst.

mod	at.001	at.01	at.10	at.20
BERT	0.0388	0.2639	0.8331	0.9116
rf	0.0908	0.3465	0.7775	0.8816
rf_cal	0.0892	0.3337	0.7682	0.8761
svm	0.1016	0.3494	0.8472	0.9346
mnb	0.0598	0.3202	0.8348	0.9333
mnb_cal	0.0273	0.2566	0.8345	0.9293



Limitations

- Analysis performed only in biological categories
  - Difficult to infer to other category type
- Fair intermediate Cohen's kappa score: instructions maybe not clear, subjectivity for classification, organisation with different perspective
  - Big set of articles to review at WHO, almost perfect agreement
- Noise definition very specific
  - Possible underestimation of the noise, but sensitivity kept
- No multivariable analysis
  - No identification of confounding or modification factors
- No assessment of the quality of the classification, no misclassification evaluation

Conclusion and recommendations (1)

- Categories and languages
  - Clean the highly correlated categories
  - Revise the categories with more than 10% of noise
  - Prioritize the revision of the key words: Turkish, Farsi and English first
  - Perform a misclassification evaluation
- Sources
  - Clean the source with more than 50% of noise
  - Rank the sources according to the proportion of noise and monitor the performance of the most important ones
- Plan to re-evaluate the proportion of noise when mitigation measures implemented to show improvement

Conclusion and recommendations (2)

For using a model as a filter (before the reports reach the users):

- Good performance is in reach, but it’s basically impossible to achieve zero false positives (the same is probably true for humans, too!).
- Classic models like MNB and SVM perform really well.
- Large-scale contextual language models like BERT may not be worth the effort.

For using a model as an interactive tool (so users can sort and filter on their own):

- The calibrated models (SVM, RF, and MNB) had the lowest Brier scores, and they produced intuitive probability distributions over the two classes of reports.
- BERT, on the other hand, was almost 100% confident in all of its guesses, whether they were right or wrong, limiting its usefulness as an interactive filtering tool.

Acknowledgments

- This work is a collaboration between WHO and US CDC
- The Joint Research Center colleagues
- The experts involved in reviewing the articles: Thomas Mollet, Jordi Borell, Cristina Rojo, Alexander Rosewell, Bryan Kim, Abdelhamid Zaghool, Paolo Tizzani, Christie Hercik, Alastair Donachie
- The EIOS Core Team members involved in this study: Elvis Chem, Johannes Schntizler, Philip Abdelmalik, Philippe Barboza
- Chad Heilig, Karen Wong, and Nirma Bustamante from US CDC

Scott Lee, PhD • Statistician • yle4@cdc.gov
Machine Intelligence and Data Science Team (MIDST)
Center for Surveillance, Epidemiology, and Laboratory Services (CSELS)

EIOS

EPIDEMIC INTELLIGENCE
FROM OPEN SOURCES

Emilie Peron, Epidemiologist, EIOS Core Team

Perone@who.int

EIOS@who.int

EXTRA SLIDES

ML Models. Use word order and context?

- **Bag-of-words** (BoW) models treat reports as unordered collections of words and phrases. Ignore order and context.
  In this study, we tried a random forest (RF), a support vector machine (SVM), and a multinomial Naïve Bayes (MNB) classifier.
- **Contextual language models** model both the meaning and the sequence of words in text and are generally more powerful than BoW models. Use order and context.
  In this study, we tried the Bidirectional Encoder Representations from Transformers (BERT) model from Google Research.

Noise proportion for the 20 most common source countries