The Potential for Misleading Results in Text Sanitisation with Standard Evaluation Metrics

Dan Zhang[1] **and Mark Anderson**[2]
[1]Norwegian University of Science and Technology
[2]Norwegian Computing Center
dan.zhang@ntnu.no, anderson@nr.no

Abstract

Data privacy is an important facet of modern life. It is especially important when considering data that carries potentially sensitive information such as in medical or legal documents. However, it is particularly difficult to ensure private information has been removed or masked in unstructured data, e.g. free-flowing text. The evaluation of systems that automatically detect and remove personal identifiable information (PII) from text is also challenging. Here we present a case study of a system that seemingly performed well, but under closer scrutiny the high performance was due to the shortcomings of standard binary classification metrics in the context of high target class prevalence. We then give a short analysis of different possible metrics in these high-prevalence scenarios, clearly showing the superiority of the Matthews Correlation Coefficient. This is particularly important because readily available data in this domain is rare and often systems are compared using biographies from Wikipedia which have a naturally high prevalence. This can be further aggravated by certain reasonable pre-processing or evaluation formalisms as in the case study discussed here.

Keywords: data sanitisation, privacy, evaluation

1. Introduction

Attempting to reduce the amount of sensitive information contained in data is not only a difficult task to undertake with textual data, but it is also not trivial to evaluate. This is partly due to standard issues of evaluating span-based tasks, such as named-entity recognition, but also due to the difficulty of having to evaluate both direct identifiers (i.e. spans that contain direct references to sensitive information that should be removed) and quasi identifiers (i.e. spans that can be used in conjunction with others to leak sensitive information). In addition, there is the question of how much of a given span needs to be detected for the risk to be considered suitably reduced.

There is also an issue with data availability, as the data of interest is sensitive by definition and cannot be readily shared. This means researcher need to use proxy datasets that carry similar characteristics to the data that actually needs to be sanitised. It is common practice to use biographies of famous individuals from Wikipedia for developing and evaluating text sanitisation systems, as suitable open-source data is difficult to develop due to privacy issues (Papadopoulou et al., 2022b; Hassan et al., 2023; Olstad et al., 2023). This Wikipedia data naturally contain a high prevalence of personal identifiable information (PII) and is particularly sensitive to the issues that will be discussed in this paper.

In situations where the prevalence of the target class is extremely high, brute-force or random methods can give results which look impressive with many conventional binary classification metrics. It is also a fairly common practice to narrow the frame of reference when undertaking research in NLP, e.g. only looking at noun phrases when attempting to detect sensitive information, as PII typically appears in noun phrases (Hassan et al., 2023; Iwendi et al., 2020). This has the effect of artificially increasing the prevalence of the target if the heuristic doesn't remove any spans/tokens that are of the target class and/or the subsequent evaluation only considers the subset of spans/tokens left after applying the heuristic. In a way, the more sound the heuristic, the more problematic it is with regards to these problems associated with conventional binary classification metrics. Further, in the space of data sanitisation, it is becoming more common to include quasi-identifiers such that the density of the text which can be considered sensitive increases.

While part of the discussion here pertains to a paper where the prevalence is artificially high due to pre-processing steps and shows the need to carefully formulate the problem space when undertaking research on free text, it is also of general interest in how practitioners should consider undertaking evaluation in the space of data sanitisation moving forward and whether using datasets derived from Wikipedia biographies (or similarly developed datasets) is suitable or not.

In this paper, we briefly introduce a case study in data sanitisation where pre-processing steps artificially increase the prevalence of sensitive spans in the subset of spans used for evaluation as an example of how the formulation of your evaluation setup can have major impacts on the validity of the evaluation. We further develop this analysis

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4638–4646
11–16 May 2026.

by evaluating a slew of different metrics used for binary classification when prevalence of the target class is particularly high, showing that most of these are not sensitive to the fact that the predictions are randomly based on the prevalence. We thus argue that the necessity of using biography-based corpora such as from Wikipedia and the often reasonable step of narrowing the frame of evaluation means that the likelihood of having to deal with a situation not dissimilar to the paper evaluated is quite high: the results presented here therefore offer some guidance to practitioners working in text sanitisation going forward when it comes to best practices for evaluation.

2. Related Work

Automated approaches for detecting sensitive information in unstructured text primarily fall into three categories: rule-based, machine learning, and hybrid systems. Rule-based methods rely on predefined patterns and dictionaries and are particularly effective for identifying structured PII such as names, dates, and locations (Neamatullah et al., 2008; Yang and Garibaldi, 2015). Machine learning models typically frame the task as sequence labelling, often leveraging named entity recognition (NER) to classify sensitive tokens (Devi and Saharia, 2025; Ahmed et al., 2021; Kleinberg et al., 2022; Pilán et al., 2022). Hybrid systems, which combine rule-based logic with machine learning, have demonstrated strong performance across diverse domains (Iwendi et al., 2020; Papadopoulou et al., 2022b).

An issue with earlier work was their focus on specific semantic categories strongly tied to NER (i.e. proper nouns, location, and so on) which resulted in certain less direct identifiers being found, as the task was setup in such a way as to not even consider them (Deleger et al., 2013; Liu et al., 2017; Dernoncourt et al., 2017; Jensen et al., 2021). However, not all named entities are identifying spans and plenty of other span types can be direct or quasi identifiers (Papadopoulou et al., 2022b). In more recent work, researchers have also attempted to include quasi identifiers in their sanitisation systems to account for this (Papadopoulou et al., 2022a,b; Pilán et al., 2022). This results in a larger portion of documents being considered to have a risk of identifying individuals. Alongside this necessary development, Weitzenboeck et al. (2022) also highlighted that unstructured data, as in text, cannot be truly de-identified (hence the use of sanitisation) because even seemingly inconspicuous spans can carry a threat of identifying individuals due to the diverse nature of language, especially as the length of text increases.

Evaluation metrics are essential for assessing the performance of sensitive information detection models. They offer quantitative insights into model behaviour and guide improvements aligned with privacy protection goals. Commonly used metrics include **precision**, **recall**, and **F1 score**: precision measures the proportion of correctly identified sensitive entities among all predictions, recall captures the proportion of actual sensitive entities detected, and F1 score balances both. Depending on the task, evaluation may be performed at the *token-level* or *entity-level*, with the latter offering stricter criteria based on complete entity spans. Some studies also introduce metrics tailored for privacy protection by applying weighted scores to direct identifiers (e.g., names, social security numbers) and quasi-identifiers (e.g., age, ZIP code), reflecting their varying disclosure risks (Pilán et al., 2022). However, these metrics remain grounded in traditional precision, recall, and F1 formulations.

Most existing methods rely on supervised learning, which requires large volumes of annotated data and often restricts detection to predefined entity types. To address these limitations, recent work has begun exploring unsupervised and zero-shot approaches aimed at identifying any potentially sensitive content without relying on annotated datasets. For instance, Hassan et al. (2023) proposed an unsupervised method using word embeddings to detect sensitive terms without employing NER. Similarly, Liu et al. (2023) introduced a zero-shot text sanitisation technique using large language models (LLMs) to estimate word-level privacy risk based on probability thresholds. In Albanese et al. (2023), ChatGPT/GPT-4 was applied to data de-identification using zero-shot and few-shot learning, with a focus on designing high-quality prompts to enhance privacy-aware performance. Papadopoulou et al. (2022a) also developed a system that did not require labelled data by leveraging knowledge graphs and an inverted index to automatically label spans in documents.

It is worth noting that these models aim to detect any form of sensitive information and are typically evaluated against NER-based baselines. However, NER models often suffer from low recall due to their reliance on predefined categories. As a result, general-purpose models for sensitive information detection frequently outperform NER baselines—even when evaluated using F_1 scores, which balance precision and recall — particularly in datasets with high prevalence of sensitive content such as biographies of known individuals taken from Wikipedia.

This raises the question: do these models truly perform better? In this paper, we use the word embedding method from Hassan et al. (2023) as a case study to demonstrate that standard metrics may produce misleading results. We further

show that the **Matthews Correlation Coefficient (MCC)**, derived directly from the confusion matrix, offers a more reliable metric for comparing general-purpose models with NER-based baselines (Matthews, 1975). Further, MCC has been considered a better option for binary classification in other areas of research (Halimu et al., 2019; Chicco and Jurman, 2023).

3. A Case Study of Misleading Results

In their paper, Hassan et al. (2023) proposed a novel method for detecting sensitive terms in textual documents using word embedding models, instead of traditional NER-based techniques. Their method and the subsequent misleading results are presented in this section as a case study.

3.1. Word embedding method and the results

The general idea is as follows. Given a document D and an entity e from D to be protected, the method first preprocesses the document D to extract noun phrases. A word embedding model (either trained or pre-trained) is used to generate vector representations of these phrases. The cosine similarity between each phrase and the protected entity e is then computed, and phrases with a similarity above a threshold are identified as sensitive terms.

Their word embedding model (Word2Vec) was trained on a corpus of abstracts from 19,000 Wikipedia articles about movie actors. For evaluation, they randomly selected 50 abstracts and manually annotated them to identify words or n-grams that could reveal the actor's identity. Annotation guidelines and further details are provided in Hassan et al. (2023).

To assess the performance of the sensitive term detection method, the paper uses precision, recall, and F_1-score. They define precision as:

$$Precison = \frac{\#\textit{detected tagged terms}}{\#\textit{detected terms}}$$

where *detected terms* are noun phrases identified by the method, and *detected tagged terms* are those containing manually tagged sensitive words. They define recall as:

$$Recall = \frac{\#\textit{detected tagged terms}}{\#\textit{tagged terms}}$$

where *tagged terms* are all manually annotated sensitive phrases.. Finally, the F_1-score is defined as the harmonic mean of precision and recall.

Table 1 compares the proposed method with several NER-based tools, including Stanford's NER3, NER4, NER7, and Microsoft's Presidio. Although the proposed approach shows lower precision, it achieves significantly higher recall, likely due to its ability to capture a wider range of sensitive terms beyond predefined categories.

This trade-off suggests improved suitability for privacy-sensitive tasks, with the F_1-score indicating overall stronger performance. However, whether these metrics truly reflect the method's effectiveness remains uncertain. The subsequent sections examine the data and methodology in greater depth to address this question.

3.2. Observations

The results from the previous section revealed that the suggested method based on cosine similarity was less effective than it initially seemed, as the cosine similarity scores between protected entities and noun phrases within the same document were generally above the threshold, but only slightly higher than scores across different documents. This suggests that cosine similarity alone may not reliably identify document-specific sensitive information, especially when scores between protected entities and many random noun phrases exceed the threshold. Longer noun phrases also tend to produce higher scores, emphasizing the influence of pre-processing.

To better understand the behaviour of the method, we considered two metric. First is **Detected Proportion** (DP):

$$Detected\_proportion = \frac{\#\textit{detected terms}}{\#\textit{all noun terms}}.$$

And second is **Prevalence**, which measures the proportion of tagged terms in the test set:

$$Prevalence = \frac{TP + FN}{N} = \frac{\#\textit{tagged terms}}{\#\textit{all noun terms}}.$$

From Table 2, unlike other methods that detect only a small fraction of noun terms as sensitive, the word embedding method identifies 96% of noun phrases after pre-processing. This aligns with earlier observations that most cosine similarity scores exceed the threshold. However, its precision matches the prevalence closely, raising concerns about its effectiveness. A naive strategy that masks all noun phrases would yield similar metrics—precision equal to prevalence, recall of 1, and an inflated F_1 score of 90.5%, outperforming all reported results. Yet, despite detecting 96% of noun phrases, the word embedding model achieves only 81.24% recall (Table 1), prompting deeper analysis.

3.3. Pre-processing

The pre-processing pipeline applies tokenization, POS tagging, and chunking to extract noun, verb,

Model	Precision	Recall	F_1
NER3	96.09%	19.59%	32.07%
NER4	97.59%	34.25%	49.72%
NER7	98.32%	27.89%	42.77%
Presidio	98.06%	27.07%	41.12%
Word Embedding	82.69%	81.24%	81.66%

Table 1: Average Precision, Recall and F_1-score for the 50 evaluated documents (Hassan et al., 2023)

Model	Prevalence	Detected proportion
NER3	82.68%	6.11%
NER4	81.72%	13.92%
NER7	81.98%	10.93%
Presidio	82.41%	12.69%
Word Embedding	82.65%	96.20%

Table 2: Statistics for the 50 evaluated documents. Prevalence rates are different due to pre-processing methods affecting both the total noun phrases found and those tagged as risky.

and prepositional phrases, which are then embedded as vectors using a word embedding model. For sensitive entity detection, only noun phrases within the same document are compared via cosine similarity to the target entity. To assess the impact of pre-processing, we compare the word embedding method (with chunking) to Presidio (without chunking).

Table 3 shows that chunking significantly reduces vocabulary size; Presidio's data size is nearly twice as large. Despite this, both methods exhibit similar prevalence, suggesting minimal impact from chunking. A key limitation of chunking is that phrases may contain varying numbers of labelled entities. Longer phrases with multiple sensitive entities are treated equally to shorter ones, which can distort evaluation metrics and model training by overlooking phrase complexity and importance.

Table 4 presents statistics for all phrases or terms, not just nouns. Comparing Table 3 and Table 4 reveals that focusing solely on noun phrases or terms increases prevalence but excludes over 20% of labelled sensitive entities. Specifically, noun-labelled phrases represent 79.5% of labelled phrases in the word embedding model, while noun-labelled terms account for 72.2% in Presidio. This explains why the word embedding model's recall is capped at 81.24% in Table 1, despite detecting approximately 96% of noun phrases (this is above the theoretical limit because some phrases are split in the pre-processing step resulting in some phrases being counted more than once). Even masking all noun phrases would yield a maximum recall of 79.5%, as the remaining 20.5% of sensitive phrases are non-nouns and thus missed entirely.

Model	Total	Labelled	Prevalence
Presidio	6348	5246	82.41%
Word embedding	3413	2804	82.65%

Table 3: Statistics for two models when only considering noun phrases. *Total* and *Labelled* given in number of spans.

Model	Total	Labelled	Prevalence
Presidio	14499	7261	50.08%
Word embedding	8598	3527	41.02%

Table 4: Statistics for two models when considering all terms or phrases. *Total* and *Labelled* given in number of spans.

This raises a key concern: is it justified to focus solely on nouns in sensitive text detection? While nouns carry strong semantic meaning, excluding non-noun elements inflates prevalence and may lead to over-predicting sensitivity in real-world applications. More importantly, it omits a significant portion of known sensitive entities, which is critical for effective privacy protection.

4. Evaluation metrics selection

In the previous section, we presented a case in which evaluation metrics suggested strong performance, but the method was less effective in practice. In this section, we explore how to choose suitable metrics for high prevalence datasets. We begin by reviewing standard binary classification metrics and then analyse various scenarios to identify robust metrics that reliably evaluate sensitive text detection.

4.1. Metrics for binary classification

In binary classification tasks, models are evaluated based on their ability to distinguish between two classes—typically labelled as **positive** and **negative**. Performance is assessed using a range of

metrics derived from the **confusion matrix**, which summarizes the outcomes of predictions:

- True Positives (TP): The model correctly predicts the positive class.
- True Negatives (TN): The model correctly predicts the negative class.
- False Positives (FP): The model incorrectly predicts the positive class when it is actually negative.
- False Negatives (FN): The model incorrectly predicts the negative class when it is actually positive.

From this matrix, several key evaluation metrics are computed. Table 5 provides a summary of the most commonly used metrics (Martín De Diego et al., 2021).

Metric	Definition
Accuracy (ACC)	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision (PPV)	$\frac{TP}{TP+FP}$
Recall (Sensitivity) (TPR)	$\frac{TP}{TP+FN}$
Specificity (TNR)	$\frac{TN}{TN+FP}$
Negative Predictive Value (NPV)	$\frac{TN}{TN+FN}$
F_1 Score (F_1)	$2 \cdot \frac{\text{PPV} \cdot \text{TPR}}{\text{PPV}+\text{TPR}}$
F_1^- Score (F_1^-)	$2 \cdot \frac{\text{NPV} \cdot \text{TNR}}{\text{NPV}+\text{TNR}}$
Fowlkes–Mallows index (FM)	$\sqrt{\text{PPV} \cdot \text{TPR}}$
Geometric Mean (GM)	$\sqrt{\text{TPR} \cdot \text{TNR}}$
Unified Performance Measure (UPM)	$2 \cdot \frac{F_1 \cdot F_1^-}{F_1+F_1^-}$
Balanced Accuracy (BA)	$\frac{\text{TPR}+\text{TNR}}{2}$
Bookmaker Informedness(BM)	$\text{TPR} + \text{TNR} - 1$
Markedness (MK)	$\text{PPV} + \text{NPV} - 1$
Matthews Correlation Coefficient (MCC)	$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

Table 5: Metrics based on a confusion matrix

To illustrate how different metrics can yield varying insights, consider a synthetic dataset with $N = 100$ instances and a class **prevalence** of 80%. For simplicity, we group the five models presented in Table 1 into two categories: the word embedding method and all other methods. Table 6 presents the confusion matrices and corresponding evaluation scores for two representative classifiers.

Classifier A exhibits precision and recall values close to the prevalence rate, resembling the performance of the word embedding method discussed earlier. In contrast, Classifier B demonstrates high precision but low recall, similar to the other methods in Table 1. We apply the full set of metrics introduced in the previous subsection to compare these two classifiers.

As observed in Table 6 and Table 7, certain metrics favour one classifier over the other. This raises an important question: *How should one select a fair and informative evaluation metric?* Choosing the right metric is especially critical in cases of **imbalanced datasets**—where one class significantly outnumbers the other—or **high-prevalence scenarios**, where some metrics may offer misleading impressions of model performance.

In the next subsection, we explore strategies for selecting appropriate evaluation metrics in imbalanced settings with 80% prevalence, and identify which metrics provide the most reliable assessments across different classification methods.

4.2. Metric selection

To select informative metrics for data with high prevalence, it is essential that the chosen metrics can effectively distinguish a model from random classifiers. Otherwise, a trivial model that predicts all instances as the majority class may yield deceptively high scores for precision, recall, and F_1, despite lacking genuine predictive power.

Figures 1, 2, and 3 illustrate how various evaluation metrics behave as an increasing proportion of the data is detected randomly (i.e. simply generating a number between 0 and 1 for each instance and labelling it as the target class if it is below the threshold set by the desired proportion). These visualizations highlight the limitations of certain metrics under high-prevalence conditions and underscore the importance of selecting metrics that remain robust against naive or biased prediction strategies.

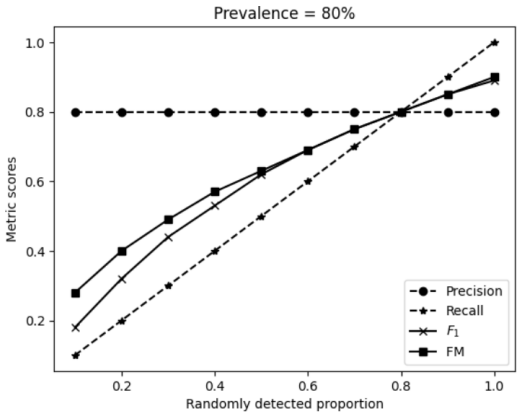

Figure 1: Precision, Recall, F_1 and FM

From Figure 1, we observe that precision, recall, F_1, and FM scores become disproportionately high when a large portion of the population is detected. This suggests that simply identifying a substantial subset of the population can inflate met-

Case	TP	FN	TN	FP	Precision	Recall	F_1
Classifier A	65	15	5	15	81.25%	81.25%	81.25%
Classifier B	24	56	19	1	96%	30%	45.71%

Table 6: Standard metrics for Classifier A and B

Case	ACC	FM	NPV	TNR	F_1^-	UPM	MCC	BA	BM	GM	MK
Classifier A	70%	81.25%	25%	25%	25%	38.23%	6.25%	53.12%	6.25%	45.06%	6.25%
Classifier B	43%	53.66%	25.33%	95%	40%	42.66%	23.09%	62.5%	25 %	53.38%	21.33%

Table 7: Other metrics for Classifier A and B

ric scores. Furthermore, as prevalence increases, these scores rise even more. In such scenarios, precision tends to converge toward the prevalence rate, while recall increases with the detected proportion. Since both F_1 and FM are derived from precision and recall, their scores also escalate. Consequently, these metrics may not reliably reflect model performance in high-prevalence settings.

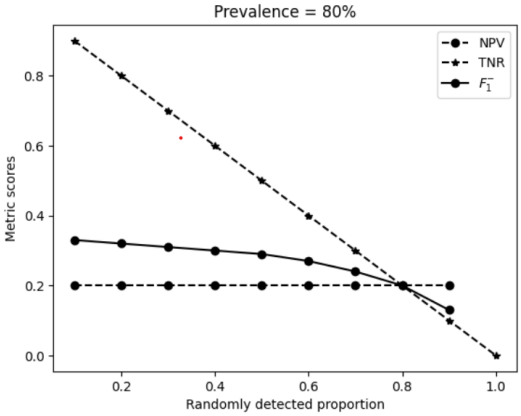

Figure 2: NPV, TNR and F_1^-

Figure 2 presents a similar analysis from the perspective of the negative class. In this context, NPV and TNR exhibit behaviour analogous to precision and recall. The metric F_1^-, defined using NPV and TNR, mirrors the behaviour of F_1 but focuses on the negative class. Precision, recall, NPV, and TNR are four basic metrics, each derived from only a subset of the confusion matrix. As a result, F_1 and FM exclude true negatives, while F_1^- omits true positives. These metrics offer partial insights into classifier performance and are individually insufficient for comprehensive evaluation.

Figure 3 introduces metrics that incorporate all four elements of the confusion matrix. Among them, MCC is directly computed from TP, TN, FP, and FN, while others are derived from the four basic metrics. UPM and GM vary with the detected proportion: their scores increase initially, peak when the detected proportion is 0.5, and then decline. In contrast, four metrics—MCC, BM, BA,

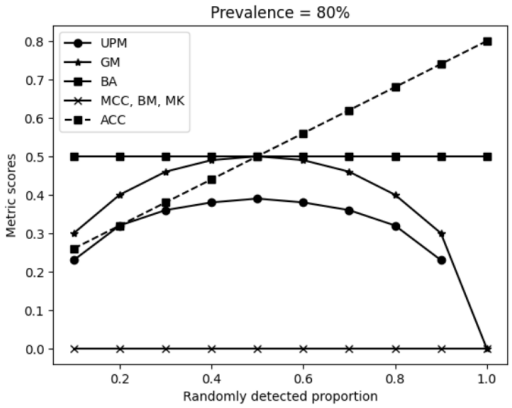

Figure 3: Other metrics

and MK—remain invariant to random detection proportions. These invariant metrics provide more robust and reliable evaluations in high-prevalence contexts. In the following subsection, we examine these four metrics in greater detail.

4.3. Comparison of MCC, BM, BA, and MK

We have identified four metrics—Matthews Correlation Coefficient (MCC), Bookmaker Informedness (BM), Balanced Accuracy (BA), and Markedness (MK)—that incorporate all four elements of the confusion matrix and are capable of detecting random classification behaviour. To determine the most suitable metric for our scenario, we conduct a deeper investigation into their properties.

Stability is a critical criterion. A reliable evaluation metric should exhibit consistent behaviour, such that small changes in the confusion matrix do not result in large, disproportionate shifts in the metric score.

Table 8 presents three illustrative cases. In Case A, the classifier predicts all instances as positive. Cases B and C each predict one instance as negative while labelling all others as positive. The key distinction is that Case B results in a false negative, whereas Case C correctly identifies a true negative.

Case	TP	FN	TN	FP	Precision	Recall	F_1	MCC	BM	BA	MK
A	80	0	0	20	0.8000	1.0000	0.8889	—	0.0000	0.5000	—
B	79	1	0	20	0.7980	0.9875	0.8826	-0.0503	-0.0125	0.4938	-0.2020
C	80	0	1	19	0.8080	1.0000	0.8938	0.2010	0.0500	0.5250	0.8081

Table 8: Metric scores for three cases with close performances

All three cases represent random or near-random classification behaviour.

Metrics such as MCC, BM, and BA successfully capture this randomness and further indicate that Case C is marginally better than Case B. However, the MK metric exhibits an abrupt and disproportionate change—jumping from -0.2 in Case B to 0.8 in Case C. This level of sensitivity is neither intuitive nor desirable when evaluating classifier performance. Based on the instability observed in Table 8, we exclude the MK metric from further consideration.

Next, we examine BM and BA using the examples in Table 9. Both metrics are derived from the True Positive Rate (TPR) and True Negative Rate (TNR), and their values remain constant as long as the sum of TPR and TNR is fixed. Consequently, all four cases in Table 9 yield identical BM and BA scores, despite clear differences in classifier performance.

Consider Cases A and B (or C and D), which represent two classifiers evaluated on the same dataset. It is evident that classifier A (C) performs better than classifier B (D). This distinction is captured by MCC and the F_1 score, but not by BM or BA.

For Cases A and C, both share the same recall, but Case C has higher precision, resulting in a higher F_1 score. However, a higher F_1 score does not necessarily imply better performance due to the higher prevalence in Case C. When precision approaches prevalence, the classifier's behaviour approximates random classification. The difference between precision and prevalence in Case A is greater than in Case C, indicating that Case A is further from random classification. Thus, despite Case C having higher precision and F_1, Case A is the superior classifier—a nuance captured only by MCC. The same reasoning applies to Cases B and D. Both have identical recall and similar precision and F_1 scores. However, Case B, with lower prevalence, is preferable to Case D. Again, MCC is the only metric that reflects this performance difference.

The above observations highlight the limitations of BM, BA, and F_1 in distinguishing classifier quality when prevalence varies. In contrast, the Matthews Correlation Coefficient (MCC) effectively incorporates prevalence and provides a more nuanced and reliable assessment of classifier performance.

5. Discussion and Conclusion

As discussed in the previous sections, pre-processing steps can alter both the size and the prevalence of the target class in a dataset. Consequently, it is essential to adopt an evaluation metric that remains robust when the formulation and subsequent evaluation space of a task leads to these asymmetric contexts. The Matthews Correlation Coefficient (MCC) satisfies this requirement, making it the most suitable metric for comparing the performance of different classification methods in this scenario. Despite other metrics ranking the two classifiers differently in Table 6 and Table 7, Classifier B outperforms Classifier A when evaluated using MCC. This reinforces the importance of selecting metrics that account for prevalence and provide a stable, comprehensive assessment of classifier quality.

While it is a known result that MCC is a more suitable metric to use when working with heavily skewed datasets, we have shown here that specifically in a text sanitisation context the use of more standard metrics such as F1, recall, and precision can lead to very misleading results and comparisons of systems. There here is a bias towards recall in text sanitisation as this captures to what extent a system performs the main task of ensuring the privacy of individuals. This then means that brute force methods such as masking every token in a document leads to an optimal solution with regards to recall and privacy. This then coupled with the second issue, that of high prevalence of risky spans in documents (or pre-processing steps and evaluation formalisms causing high prevalence), results in high precision scores as shown in section 4. This then can lead to absurd situations where brute force methods perform well compared to legitimate methods when using these metrics.

While this high prevalence is particularly apparent in Wiki-based corpora using biographies of well-known individuals due to the density of personal information contained therein, the prevalence in more applicable corpora such as the TAB corpus based on ECHR documents can still be somewhat high depending on how the evaluation is undertaken (Pilán et al., 2022). As shown in section 3, the prevalence can be artificially inflated by narrowing the evaluation space, in that case to only considering noun phrases. The TAB corpus was

Case	TP	FN	TN	FP	Prevalence	DP	Precision	Recall	F1	MCC	BA	BM
A	70	5	10	5	83.3%	83.3%	93.3%	93.3%	93.3%	60.0%	60.0%	80.0%
B	50	25	14	1	83.3%	56.7%	98.0%	66.7%	79.4%	45.1%	60.0%	80.0%
C	140	10	10	5	90.9%	87.9%	96.6%	93.3%	94.9%	52.9%	60.0%	80.0%
D	100	50	14	1	90.9%	61.2%	99.0%	66.7%	79.7%	35.4%	60.0%	80.0%

Table 9: Different performances with the same BA ad BM scores

Case	TP	FN	TN	FP	Prevalence	DP	Precision	Recall	F1	MCC	BA	BM
A	26	24	35	15	50%	41%	64.31%	52%	57.14%	22.37%	22%	61%
B	13	37	46	4	50%	17%	76.47%	26%	38.81%	23.96%	18%	59%
C	35	15	26	24	50%	59%	59.32%	70%	64.22%	22.37%	22%	61%

Table 10: Differences between evaluation metrics for balanced data

annotated to focus on spans corresponding to entities and if evaluation was restricted to whether a given entity posed a risk of identifying an individual or not, the prevalence would be 68% (not far removed from the prevalence used in our analysis in section 4). Whereas focusing on just noun phrases (as was done in (Hassan et al., 2023)) the prevalence would just be 24% and considering all phrases, it would be as low as 8%. This highlights the issue on how making seemingly reasonable decisions, such as focusing on a subtype of spans because PII rarely occur in other types, can lead to unexpected issues in evaluation. Beyond this even if the prevalence is not particularly high such that the data is imbalanced, precision can still give misleading interpretations because if precision is similar to the prevalence then this suggests that the predictions are not much better than random.

Furthermore, this situation of high prevalence could arise naturally in the future and not just through ill-considered formalisms with the density of risky spans in documents increasing due to potential for even seemingly innocuous spans carrying an identifying risk as discussed above and in detail in (Weitzenboeck et al., 2022). Therefore we advocate for the wider adoption of MCC in the text sanitisation space and for the potential development of specific variants of the metric similar to the work in (Pilán et al., 2022) where F1, precision, and recall were altered to better capture the nuances of evaluating text sanitisation.

While we have focused on the issue of evaluating models on data with a high prevalence of risky spans, it is important to note that standard metrics can also produce misleading results in more balanced datasets. As shown in Table 10, this issue is not limited to corpora with dense concentrations of identifying spans. More broadly, the F_1 score tends to suggest better performance than in reality and also results in a different ranking of systems.

We therefore advocate the use of MCC at least in conjugation with other metrics, so that seemingly harmless decisions regarding the formalism of the task undertaken or the use of heavily imbalanced data does not lead to unforeseen issues with misleading evaluation. We hope future work will look to develop MCC specifically for the use in text sanitisation or to derive complement metrics that help analysis the efficacy of a system specifically with regards to privacy prevention and utility, considering MCC is a composite of both of these concerns in the same vein as F1.

6. Bibliographical References

Hadeer Ahmed, Issa Traore, Sherif Saad, and Mohammad Mamun. 2021. Automated detection of unstructured context-dependent sensitive information using deep learning. *Internet of Things*, 16:100444.

Federico Albanese, Daniel Ciolek, and Nicolas D'Ippolito. 2023. Text sanitization beyond specific domains: Zero-shot redaction & substitution with large language models. *arXiv preprint arXiv:2311.10785*.

Davide Chicco and Giuseppe Jurman. 2023. The matthews correlation coefficient (mcc) should replace the roc auc as the standard metric for assessing binary classification. *BioData Mining*, 16(1):4.

Louise Deleger, Katalin Molnar, Guergana Savova, Fei Xia, Todd Lingren, Qi Li, Keith Marsolo, Anil Jegga, Megan Kaiser, Laura Stoutenborough, et al. 2013. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *Journal of the American Medical Informatics Association*, 20(1):84–94.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Longjam V. Devi and N. Saharia. 2025. Detection of sensitive information using crf in unstructured text. In *Recent Advances in NLP*, chapter 18. CRC Press.

Chongomweru Halimu, Asem Kasem, and SH Shah Newaz. 2019. Empirical comparison of area under roc curve (auc) and mathew correlation coefficient (mcc) for evaluating machine learning algorithms on imbalanced datasets for binary classification. In *Proceedings of the 3rd international conference on machine learning and soft computing*, pages 1–6.

Fadi Hassan, David Sánchez, and Josep Domingo-Ferrer. 2023. Utility-preserving privacy protection of textual documents via word embeddings. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):1058–1071.

Celestine Iwendi, Syed Atif Moqurrab, Adeel Anjum, Sangeen Khan, Senthilkumar Mohan, and Gautam Srivastava. 2020. N-sanitization: A semantic privacy-preserving framework for unstructured medical datasets. *Computer Communications*, 161:160–171.

Kristian Nørgaard Jensen, Mike Zhang, and Barbara Plank. 2021. De-identification of privacy-related entities in job postings. *NoDaLiDa 2021*, page 210.

Bennett Kleinberg, Toby Davies, and Maximilian Mozes. 2022. Textwash–automated open-source text anonymisation. *arXiv preprint arXiv:2208.13081*.

Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of biomedical informatics*, 75:S34–S42.

Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, et al. 2023. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*.

Isaac Martín De Diego, Ana R Redondo, Rubén R Fernández, Jorge Navarro, and Javier M Moguerza. 2021. General performance score for classification problems.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Ishna Neamatullah, Margaret M Douglass, Li-Wei H Lehman, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. 2008. Automated de-identification of free-text medical records. *BMC medical informatics and decision making*, 8(1):32.

Annika Willoch Olstad, Anthi Papadopoulou, and Pierre Lison. 2023. Generation of replacement options in text sanitization. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 292–300, Tórshavn, Faroe Islands. University of Tartu Library.

Anthi Papadopoulou, Pierre Lison, Lilja Øvrelid, and Ildikó Pilán. 2022a. Bootstrapping text anonymization models with distant supervision. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4477–4487, Marseille, France. European Language Resources Association.

Anthi Papadopoulou, Yunhao Yu, Pierre Lison, and Lilja Øvrelid. 2022b. Neural text sanitization with explicit measures of privacy risk. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 217–229.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.

Emily M Weitzenboeck, Pierre Lison, Malgorzata Cyndecka, and Malcolm Langford. 2022. The gdpr and unstructured data: is anonymization possible? *International Data Privacy Law*, 12(3):184–206.

Hui Yang and Jonathan M Garibaldi. 2015. Automatic detection of protected health information from clinic narratives. *Journal of biomedical informatics*, 58:S30–S38.