Private Sensitive Content on Social Media: An Analysis and Automated Detection for Norwegian

Haldis Borgen[1] Oline Zachariassen[2] Pelin Mişe[3] Ahmet Yıldız[3] and Özlem Özgöbek[4]

***Abstract*— This study addresses the notable gap in research on detecting private-sensitive content within Norwegian social media by creating and annotating a dataset, tailored specifically to capture the linguistic and cultural nuances of Norwegian social media discourse. Utilizing Reddit as a primary data source, entries were compiled and cleaned, resulting in a comprehensive dataset of 4482 rows. Our research methodology encompassed evaluating a variety of computational models—including machine learning, deep learning, and transformers—to assess their effectiveness in identifying sensitive content. Among these, the NB BERT-based classifier emerged as the proficient, showcasing accuracy and F-1 score. This classifier demonstrated remarkable effectiveness, achieving an accuracy of 82.75% and an F1-score of 82.39%, underscoring its adeptness at navigating the complexities of privacy-sensitive content detection in Norwegian social media. This endeavor not only paves the way for enhanced privacy-sensitive content detection in Norwegian social media but also sets a precedent for future research in the domain, emphasizing the critical role of tailored datasets in advancing the field.**

I. INTRODUCTION

The use of social media has revolutionized the way people connect online. The revolution provided easy and inexpensive means of sharing information and expressing opinions but also brought many problems related to potential violations of users' privacy [1]. In parallel, with the rapid advancement of technology and globalization, protecting personal data has become challenging. Individuals are increasingly sharing personal information publicly and globally, leading to an observed increase in the sharing and collection of private sensitive data [2]. In the literature, private sensitive content was defined in different approaches such as using the visibility or anonymity of the user posting, utilizing privacy dictionaries to search for sensitive words or terms [3], considering sensitive topics independent of personal identification [4]. Instead of various definitions, utilizing a definition aiming to align with the European General Data Protection Regulation (GDPR) was considered a means to obtaining more aligned findings. Personal data has become a valuable resource for targeted marketing, data analytics, and potentially intrusive purposes. Detecting the shared private sensitive data via social media is crucial for getting ahead of negative consequences to individuals. Users often regret what they post on social media, partly due to oversharing or reaching an unintended audience. The findings of a study showed that people may be unaware that they are posting something they will later regret, and that the reactions of others to the content contribute to the regret [5]. Another study showed that the likelihood of post-related regret and potential repercussions can be minimized by implementing a system that warns or notifies users before sharing something private or sensitive on social media [6]. In a study related to Detecting and Grading Hateful Messages in the Norwegian Language[7], a dataset that was collected from several social media platforms in Norwegian was used and notably, the most heavily debated posts from Facebook, Twitter and Resett focused on immigration, the environment, and politics. After examining the data set, it was discovered that there was a significant imbalance because a great majority contained political opinions that must be labeled as private-sensitive. Detecting and classifying sensitive and non-sensitive contents in social media can be done using machine learning and deep learning techniques.

Due to the limited number of studies exploring Norwegian social media for detecting private sensitive content, a noticeable gap exists in the availability of suitable datasets for such private sensitive data detection. This research gap underscores the necessity to collect a relevant dataset, and Reddit has been chosen as the primary data source. One of the primary reasons for selecting Reddit is its characteristic of hosting publicly available data.

The creation of a new dataset becomes important in addressing this research gap. The significance lies in establishing a foundation for studying private-sensitive content detection on Norwegian social media. Furthermore, it not only bridges the gap in research pertaining to Norwegian social media but also lays the groundwork for future investigations into private-sensitive content detection. The creation of this dataset, driven by the absence of existing studies and suitable datasets, is fundamental in advancing research in this domain.

The objective of this study is to show how to align the definition of private-sensitive content with the GDPR, processing the collected data and preparing that can be used in future works in Norwegian language. It is also aimed to use the created dataset to detect private-sensitive user-generated content on social media platforms written in the Norwegian language using machine learning, deep learning approaches and comparison of different approximations' performance.

[1]twoday, Oslo, Norway haldis.k.borgen at twoday.com
[2] PwC, Oslo, Norway olinezac at pwc.com
[3]MEF University, Department of Computer Engineering, Istanbul, Turkey misepe, yildizah at mef.edu.tr
[4]Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway ozlem.ozgobek at ntnu.no

II. Dataset Creation and Pre-processing

A. Defining Private Sensitive Content

The variability in the criteria employed for identifying private sensitive content can precipitate disparities in the detection of such content, primarily because these criteria are not always anchored in uniform principles. The inherently subjective nature of privacy complicates the establishment of a universally accepted definition of what constitutes private-sensitive content. This necessitates the development of well-defined guidelines for annotating datasets to ensure a shared understanding. In [8], the authors presented annotation guidelines for privacy sensitive content on social media for English language. Following their work, in this paper we introduce annotation guidelines specifically adapted for Norwegian social media posts which considers the specifics of the language and the principles encapsulated within the GDPR. The objective is to instill a degree of uniformity and objectivity in the annotation process by delineating the categories of content to an exhaustive extent, thereby minimizing instances where annotators might confront discomfort or ambiguity beyond the predefined categories.

The guidelines we present are specifically adapted to the milieu of Norwegian social media, integrating aspects of the Norwegian language, including prevalent slang and abbreviations, to cater to the particular linguistic and cultural context. This adaptation was important for ensuring that the process of identifying and categorizing content as private or sensitive was persistent and comprehensive within the Norwegian social media landscape. Moreover, we introduce additional guideline categories that correspond with the types of data deemed by the GDPR as warranting protection. These encompass an individual's financial status, personally identifying information (PII), and non-public data pertaining to criminal activities, among others. The incorporation of these supplementary categories reflects the GDPR's broad scope in safeguarding various facets of personal data against unauthorized exposure or dissemination.

For a Norwegian social media text to be defined as private sensitive, it has to be within at least one of the following categories [11]:

- Personally Identifiable Information (PII)
- Information about the location of the author or other individuals mentioned
- Physical or mental status
- Details about family or romantic relationships
- Information about one's economic condition
- Indications of potential political or religious inclinations of an individual
- Information about one's non-public details about illegal actions

Based on these annotation guidelines we present, the dataset has been labeled in four classes: Sensitive, non-sensitive, unknown and unintelligible.

TABLE I
Details of Data Amounts

Subreddit	# of Entries
r/norge	13427
r/oslo	7642
Merged and cleaned data set	20852
Annotated data set	4482

B. Data Collection and Annotation

The crucial part of this study has been dedicated to the data collection and annotation processes due to not having the appropriate dataset in the academic literature for detecting private sensitive content consisting of Norwegian entries from social media. Moreover, a significant aim was to contribute to the research landscape by generating a labeled dataset specific to Norwegian social media. Rather than expending additional resources on seeking out pre-existing Norwegian datasets, a new initiative was launched to establish a dataset tailored to detecting private sensitive content in the Norwegian context. Reddit[1], a social media platform facilitating content sharing and discussions, was selected as the source due to its abundance of publicly accessible data and the prevalence of informal language usage among its user base. Leveraging the Reddit API and PRAW (Python Reddit API Wrapper), the process involved delving into the r/Norge and r/Oslo subreddits to extract Norwegian content. This approach yielded a total of 21,069 entries within the dataset. Subsequently, to ensure the data quality and the integrity of subsequent analyses, the raw data underwent a rigorous cleaning phase to eliminate inconsistencies, errors, and missing values. This meticulous cleaning process resulted in a temporary dataset comprising 20,852 rows which is then reduced to a final 4482 row dataset after further cleaning and annotating. A comprehensive breakdown of the data obtained through the scraping of the r/norge and r/oslo subreddits, along with insights into the merged and refined dataset, can be seen in the Table I. [2]

A total of eight volunteer annotators participated in the annotation of the final annotated dataset. In the annotation process, each entry was annotated by at least two annotators, with a target of involving three annotators whenever possible.

For a text to be labeled as sensitive, annotators were instructed to consider not just the explicit mention of sensitive categories but also the context in which information was presented. For instance, the mention of medical conditions or medications was considered sensitive, especially when linked to an identifiable individual. Similarly, financial information, even if it appeared benign or generic, was classified as sensitive if it could impact an individual's privacy or financial security. This process required annotators to engage in critical thinking and sometimes discussions with other peers.

[1]https://www.reddit.com

[2]The final annotated dataset with 4482 entries is available for research purposes and can be requested here: https://github.com/haldisborgen/Detecting-private-sensitive-content-in-Norwegian-Social-Media

TABLE II
AGREEMENT STATISTICS AMONG ANNOTATORS

Cluster number	Cohen's kappa	Fleiss' kappa	At least two annot. agree	All three annot. agree
Cluster 1	-	0.245416	92%	41.1%
Cluster 2	0.834513	-	88%	-
Cluster 3	-	0.764783	98.5%	80.7%
Cluster 4	0.865853	-	93.2%	-
Cluster 5	0.846013	-	89.8%	-

We specifically avoided the discussions among annotators to avoid bias.

When determining if information was non-sensitive, annotators looked for content that discussed general topics, shared widely known facts, or involved impersonal dialogue. The guiding principle was whether the text could reasonably be expected to infringe on someone's privacy or lead to identification. Information deemed public knowledge, like comments on public figures or events, was typically labeled non-sensitive, provided it did not cross into personal opinions or information about the poster or others that could be deemed private.

The unknown category was reserved for instances where context or content did not provide enough clarity to make an informed decision. This often applied to texts with vague references, lacking explicit mentions of sensitive information or clear non-sensitive content. Annotators were encouraged to use this category sparingly, aiming to resolve uncertainties through research or consultation with peers. However, when ambiguity remained despite these efforts, labeling content as unknown ensured that potentially sensitive information was not mistakenly categorized as non-sensitive.

To distribute the data to be labeled among the annotators, we have splitted the dataset into five clusters which are unequal in size. Two of these clusters (Cluster 1 and Cluster 3) have been annotated by three annotators, and three of them (Cluster 2, Cluster 4, Cluster 5) by two annotators where each of these clusters constitute approximately half of the total data. Ideally, each data set would be labeled by three different annotators, but due to limited resources, some of the data sets were labeled by two annotators. After receiving all the annotations, we have looked into the annotator agreement statistics.

To provide insight into the agreement among annotators, statistics were calculated for annotation agreement using the metrics Fleiss' kappa and Cohen's kappa. These statistics were calculated based on the number of annotators involved in each of the five clusters. Fleiss' kappa is used to calculate the annotator agreement for the clusters labeled by three annotators and Cohen's kappa is used for the clusters labeled by two annotators. Additionally, the percentage of the annotation results where at least two or all three annotators agree were computed. Table II shows the statistics of annotator agreements.

The level of agreement among annotators can provide insights into various aspects, such as the annotation guidelines' effectiveness, the annotators' precision, or the annotation task's difficulty. For the "Sensitive" class, the majority of annotations were assigned when only one of the annotators labeled it as private-sensitive. This count decreases from 228 to 145 when exactly two annotators agree on the label, and further decreases to 88 when all three annotators agree. A similar trend is observed for the "Unknown" label. In contrast, the "Non-sensitive" and "Unintelligible" classes have a majority when all annotators agree.

C. Data Cleaning and Pre-processing

A recurrent error was the classification of text as belonging to a private-sensitive subcategory while simultaneously being labeled as unknown, unintelligible, or non-sensitive. It was imperative to ensure that every entry marked as private-sensitive also carried the classification of at least one relevant subcategory, and vice versa. In order to facilitate model training, all datasets were amalgamated into a single entity referred to as the merged dataset. To enhance the feasibility of analysis, the post title, designated as "title," and the content of a post/comment, denoted as "selftext," were combined. If a post contained text in both columns, the title was positioned at the commencement of the selftext column, followed by the original selftext content. Some entries possessed a title but lacked text content.

For the experiment, various pre-processing steps were performed on the data set. Firstly, all columns except "content," "non-sensitive," "unintelligible," "unknown," and "sensitive" were removed. The "unknown" and "unintelligible" columns were merged into a single column called "other." Furthermore, any NaN values were replaced with empty strings. Certain characters, such as ']', '[', '(', and ')', were removed, and all URLs were replaced with the placeholder "@LINK". Finally, all labels were combined into a single column called "Label," which can contain one of the three labels for each entry. The resulting processed data set consists of 4,442 entries, with 981 labeled as "sensitive," 1,693 labeled as non-sensitive, and 1,768 labeled as unknown. To analyze the distribution to better understand any lexicographic patterns it was crucial to consider text length. To further understand the relationship between text length and the different classes distributions can be seen in figure 1.

In Figure 2, it was evident that the data set primarily comprises shorter texts, with a significant majority falling into this category. However, there were also outliers, represented by a few instances that contain over 800 words.

A few outliers were observed, manifested in instances containing over 800 words. However, due to their limited presence, this aspect wasn't deemed critical within the preprocessing stage. Following annotation by the annotators, the collected dataset underwent pre-processing to be made ready for model training. The annotated dataset was then partitioned into a training set and a test set. The training set was deliberately balanced, encompassing 940 entries for each of the three categories. This balanced distribution was strategically chosen to ensure equitable representation during the model training process. The training set, as a whole, comprises 2,820 entries. In contrast, the test set was

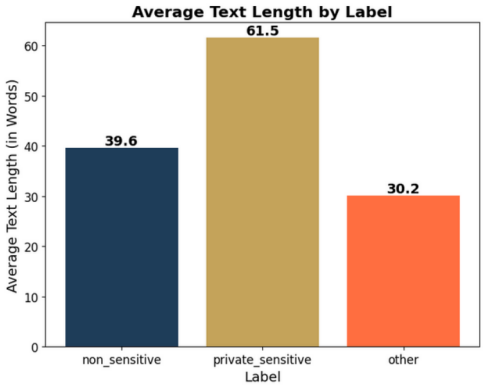

Fig. 1. Average text length in each class

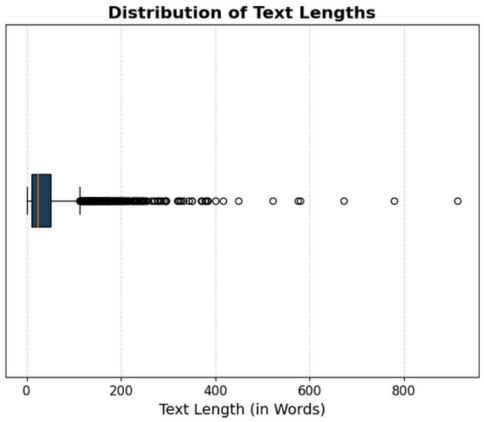

Fig. 2. Text length distribution

intentionally rendered unbalanced. The objective here was to mimic the category distribution observed in the real-world domain being represented. Consequently, the test set adopts an approximate 90/10 distribution. This equates to 49 entries for the private-sensitive category, 115 entries for the non-sensitive category, and 120 entries for the other category. The test set totals 284 entries.

III. METHODS

The research methodology adopted in this study is designed to meticulously evaluate the efficacy of various computational approaches, specifically machine learning, deep learning, and transformers, in the context of identifying private and sensitive content on Norwegian social media platforms. This comprehensive approach is essential for understanding the nuances and complexities associated with the automated detection of such content, given the unique linguistic and cultural characteristics of the Norwegian social media landscape.

The evaluation of these diverse approaches provides a foundation for assessing the effectiveness of machine learning, deep learning, and transformer models in identifying private-sensitive content within the specific domain of Norwegian social media. The research methodology emphasizes the importance of adapting models to the linguistic characteristics of social media language while maintaining an evaluation process to ensure generalizability and performance on unseen data.

A. Machine Learning

Machine learning techniques, the first of the approaches under consideration, involve training algorithms to classify content based on features extracted from the data. This method relies on the identification of patterns and characteristics within the data that are indicative of private-sensitive content. The process includes the selection of relevant features, training the model on a subset of the data, and then testing its ability to accurately identify sensitive content in an unseen dataset. This approach is critical for establishing a baseline for performance and understanding the limitations and capabilities of more traditional computational models in detecting sensitive content.

Following the pre-processing, the dataset was streamlined to encompass the label column and the content column. In this study, four distinct machine learning algorithms were employed to address the task of detecting private sensitive content. The convolutional classifiers utilized were:

- Multinomial Logistic Regression (Multinomial LR)
- Multinomial Naive Bayes
- Random Forest
- Linear Support Vector Machine (Linear SVM)

Upon training the machine learning models, their outputs were dedicated to discerning whether the content of a given text contained sensitive information or not. In cases where such a determination couldn't be confidently made, the prediction was categorized as "unknown." The workflow commenced with hyperparameter tuning via grid search, aimed at identifying the optimal hyperparameters for each model. The selected hyperparameters were determined from the grid search results for each machine learning approaches. Capitalizing on these optimal hyperparameters, the models were trained to effectively detect private sensitive content within Norwegian social media.

B. Deep Learning

Deep learning, the second approach, represents an advancement over traditional machine learning techniques by employing neural networks with multiple layers. These models are capable of capturing more complex patterns in the data by automatically discovering the representations needed for classification from raw data. In the realm of detecting private and sensitive content within social media posts, deep learning models have significance. For this study three different deep learning algorithms were used to detect private sensitive content in Norwegian Social Media. The employed deep learning models are as follows:

- Long Short-Term Memory (LSTM)
- Bidirectional Long Short-Term Memory (BiLSTM)
- Gated Recurrent Units (GRU)

The introduction of Bidirectional LSTM takes the analysis a step further by leveraging not only the preceding context

but also the subsequent context of each word or token. This holistic understanding is essential for accurate detection of private content, as the relevance of certain information might be influenced by what comes both before and after a specific phrase. BiLSTM's dual perspective helps capture nuanced contextual cues that may be missed by unidirectional models. Gated Recurrent Units offer a compelling alternative to LSTM, combining memory efficiency with similar modeling capabilities. For detecting sensitive content, GRU aids in effectively capturing the temporal patterns within social media data, allowing the model to recognize recurring themes, keywords, or expressions that might signify private or confidential information. Its simplified architecture also contributes to faster training and prediction times, which can be crucial for real-time content analysis on social media platforms.

Different activation functions, dropout values and optimizers were used to find the most efficient model for each three algorithms. Grid search approximation was applied to see all possible combinations of the activation functions, dropouts and optimizers. Findings based on grid search were used to decide which hyperparameters can be chosen for the best of each algorithm. According to grid search three different models were trained with the same dataset that was collected in this study.

Given the rapid pace at which social media content is generated, the deployment of these advanced neural architectures is instrumental in upholding privacy and security standards across digital platforms.

C. Transformers

Transformers, the third approach, introduce an even more sophisticated mechanism for modeling relationships in data. Transformers utilize self-attention mechanisms to weigh the significance of different parts of the input data differently, allowing for a more nuanced understanding of context and the relationships between words or features in a dataset. BERT is a transformer model which stands for Bidirectional Encoder Representations from Transformers [9], [10]. NB BERT-base is a specific BERT-base model that draws its training from the extensive digital archive of the National Library of Norway. This model mirrors the architecture of the BERT Cased multilingual model while being fine-tuned on a diverse range of texts in the Norwegian language. To improve the performance of the NB-BERT Base model to account for the distinctive nuances and characteristics of Norwegian social media language, and make the most of the available labeled and unlabeled data, it employed a masked language model as the domain adaptation technique.

By employing domain adaptation, the proposed approach aims to enhance the NB-BERT Base model's effectiveness in accurately categorizing private-sensitive content within the specific domain of Norwegian social media. The model's performance is gauged in each iteration, and the outcomes are synthesized by computing the average performance across all folds. To identify the most optimal parameter combination, grid search is conducted through repeated cross-validation. For the Bert model, this procedure includes the implementation of early stopping. This comprehensive approach capitalizes on a significant portion of the data for training the final model, all the while maintaining a rigorous evaluation process through K-fold cross-validation. This strategy bolsters confidence in the model's generalizability and performance on previously unseen data.

The identical dataset employed to construct a model for detecting private-sensitive content in Norwegian social media is also utilized here. Various combinations of hyperparameters are explored through a grid search integrated with cross-validation. The key parameters under scrutiny are the learning rate and the number of epochs. Additionally, early stopping is incorporated to ascertain the optimal number of epochs by considering the diverse learning rates experimented with. The optimal learning rate and number of epochs are determined based on the outcomes from the grid search and the evaluations performed during the early stopping phase.

The collected dataset serves as the foundation for assessing the performance of machine learning, deep learning, and transformer-based models in the task of identifying private-sensitive content in Norwegian social media. In evaluating model performance during the grid search, the mean accuracy across all folds is computed. Moreover, for each epoch within the cross-validation process, both validation loss and training loss are calculated. These metrics contribute to generating a graphical representation illustrating the average validation loss and training loss across all folds in the cross-validation process. [3]

IV. RESULTS

Different approaches which are conventional classifiers, deep learning and transformer based models were employed to evaluate the performance of detecting private sensitive content based on the unbalanced test dataset.

A. Machine Learning Results

Within the realm of conventional classifiers, the study selected a suite of algorithms known for their robustness and versatility in various machine learning tasks. These included Multinomial Logistic Regression (LR), Multinomial Naive Bayes (NB), Random Forest (RF), and Linear Support Vector Machine (SVM). Each of these classifiers brings a distinct set of strengths and computational strategies to the task, offering a broad perspective on the potential for traditional machine learning techniques in the realm of sensitive content detection. Considering overall performance on the test set and specifically for the private-sensitive class, Multinomial LR emerged as the better-performing classifier across various metrics, followed by Random Forest. As it can be seen in Table III, Multinomial LR attained the highest overall F1-score and precision.

[3]The code and dataset is available for research purposes and can be requested here: https://github.com/haldisborgen/Detecting-private-sensitive-content-in-Norwegian-Social-Media

TABLE III

PERFORMANCE EVALUATION ON THE TEST SET OBTAINED FROM THE CONVENTIONAL CLASSIFIERS.

Classifier	Accuracy	F1-score	Precision	Recall
Multinomial LR	**0.7430**	**0.7290**	**0.7228**	**0.7637**
Multinomial Naive Bayes	0.5563	0.5383	0.6610	0.6312
Random Forest	0.7218	0.6988	0.7026	0.7296
Linear SVM	0.6373	0.6086	0.6579	0.6806

Figures 3 and 4, display the confusion matrices, outlining predicted versus true labels for conventional classifiers on the test dataset with the chosen hyperparameter combination. These matrices illustrate the correct and incorrect classifications for each label category, offering valuable insights into classifier performance across different labels. Diagonal elements represent accurately classified samples, while off-diagonal elements signify misclassifications.

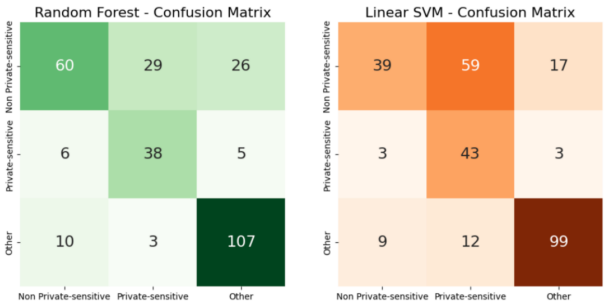

Fig. 3. Confusion matrices for Random Forest and Linear SVM respectively

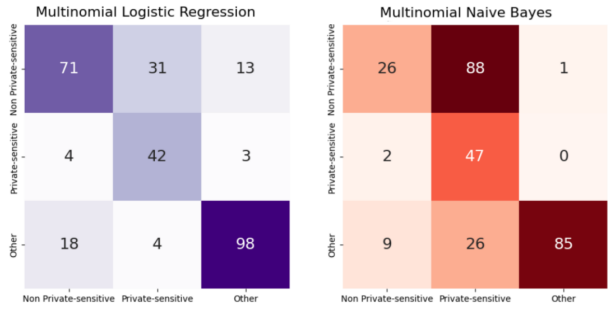

Fig. 4. Confusion matrices for Multinomial LR and Naive Bayes respectively

B. Deep Learning Results

After conducting an extensive grid search to optimize the hyperparameters for each approach, including BiLSTM, GRU, and LSTM, the best configurations for dropouts, activation functions, and optimizers were meticulously determined. This process involved a systematic exploration of various combinations of these parameters to identify the most effective setup for each neural network architecture. The outcomes of these experiments are detailed in the Table IV, showcasing the results obtained through deep learning methodologies.

TABLE IV

PERFORMANCE EVALUATION ON THE TEST SET OBTAINED FROM THE DEEP LEARNING ALGORITHMS

Classifier	Accuracy	F1-score	Precision	Recall
LSTM	0.7148	0.7235	0.7518	0.7147
BiLSTM	0.7077	0.7224	0.7077	0.7066
GRU	0.7147	0.7168	0.7330	0.7147

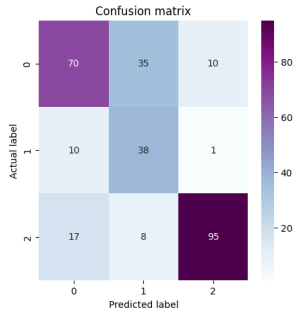

Fig. 5. Confusion Matrix For LSTM

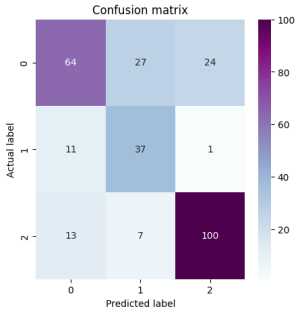

Fig. 6. Confusion Matrix For BiLSTM

C. NB BERT-based Classifier Results

With considering the grid search, appropriate epoch and learning rate values were determined. Table V displays the outcomes of the fine-tuned NB BERT-based models on the unbalanced test set. Notably, the most successful model emerged from training with 2 epochs and a learning rate of 10^{-5} .

TABLE V

PERFORMANCE EVALUATION ON THE NB BERT-BASED MODELS WITH DIFFERENT LEARNING RATES AND NUMBERS OF EPOCHS.

Learning Rate	Epoch	Accuracy	F1-score	Precision	Recall
1e-6	6	0.8028	0.7969	0.7895	0.8292
1e-5	**2**	**0.8275**	**0.8239**	**0.8106**	**0.8525**

In the Figure 8, private sensitive content detection using NB-BERT model can be seen. These findings indicate that

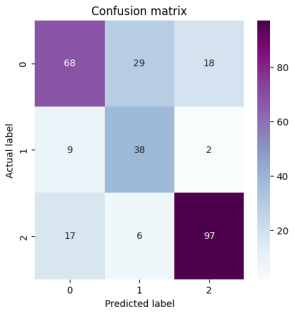

Fig. 7. Confusion Matrix For GRU

the model excels at identifying and accurately categorizing private-sensitive cases. This is supported by the confusion matrix, which shows that only 2 private-sensitive instances were mislabeled.

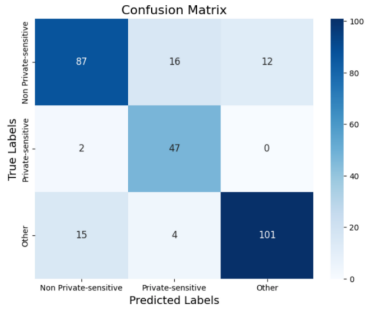

Fig. 8. Confusion matrix for NB-Bert Classifier

V. DISCUSSION

The study presents a comprehensive evaluation of various AI models in detecting private-sensitive content within Norwegian social media data, employing machine learning, deep learning, and transformer approaches. The results showcase distinct performance characteristics of each model type.

The conventional classifiers, including Multinomial Logistic Regression (Multinomial LR), Multinomial Naive Bayes, Random Forest, and Linear Support Vector Machine (Linear SVM), demonstrated varying degrees of effectiveness. Multinomial LR emerged as the most balanced model, achieving the highest overall F1-score and precision. This suggests that logistic regression, with its linear decision boundaries, is particularly adept at handling the nuances of text classification in this context. The Random Forest algorithm also showed commendable performance, indicating that ensemble methods can effectively capture the complexity of textual data. However, the Naive Bayes and Linear SVM, despite their higher recall, fell short in precision, which could be attributed to their probabilistic and margin-based decision principles, respectively.

In the deep learning domain, LSTM, BiLSTM, and GRU were evaluated. LSTM and BiLSTM models showed competitive performance, with BiLSTM slightly lagging behind in recall. This indicates that capturing both past and future context in BiLSTM does not significantly enhance performance for this specific task, possibly due to the nature of the data where context in either direction is equally informative. GRU, with its simpler structure, performed comparably to LSTM, demonstrating its efficiency in capturing temporal patterns in text data.

The transformer model, specifically the NB BERT-based classifier, outperformed both machine learning and deep learning models in terms of accuracy, F1-score, precision, and recall. This superior performance can be attributed to BERT's deep bidirectional nature, which allows for a more nuanced understanding of context and language semantics. The domain adaptation technique further refined its capabilities for the Norwegian language context, making it exceptionally adept at identifying subtle indicators of private-sensitive content.

The study's findings underscore the importance of model selection based on the nature of the data and the task at hand. In detecting private-sensitive content:

Among traditional machine learning approaches, Multinomial Logistic Regression demonstrated the most effective performance, with an accuracy of 74.30% and an F1-score of 72.90%. This model's strength lies in its precision and recall, making it an excellent choice for scenarios where a balance between computational efficiency and accuracy is required. Random Forest also performed commendably but lagged slightly behind Logistic Regression in terms of precision and recall. Other models like Multinomial Naive Bayes and Linear SVM showed a tendency to overclassify texts as sensitive, reflected in their lower precision scores.

Deep learning models, specifically LSTM, BiLSTM, and GRU, showcased their prowess in handling complex language structures. LSTM stood out with an accuracy of 71.48% and an F1-score of 72.35%, indicating its efficiency in sequential data processing. BiLSTM, though offering the advantage of understanding both past and future contexts, did not significantly surpass LSTM in this specific task. GRU, with a simpler architecture, closely matched LSTM's performance, making it a viable alternative where computational resources are limited.

The NB BERT-based classifier emerged as the most proficient model, with an outstanding accuracy of 82.75% and an F1-score of 82.39%. Its superior performance is attributed to its deep bidirectional nature and the ability to grasp nuanced contextual information, crucial for accurately identifying private-sensitive content. The model's high precision and recall indicate its exceptional capability in both correctly identifying sensitive content and minimizing false positives.

VI. CONCLUSION & FUTURE WORK

The choice of the most suitable AI model for detecting private-sensitive content depends on specific requirements such as desired accuracy, computational resources, and the nature of the dataset. The NB BERT-based classifier is recommended for tasks where high accuracy and comprehensive detection are paramount. In contrast, for contexts where

computational efficiency is a concern, models like Multinomial Logistic Regression and LSTM provide a balanced solution. Each model has its strengths, and the selection should align with the task's objectives and constraints. This analysis provides a roadmap for selecting the appropriate model based on performance metrics and application needs.

Future studies in the area of social media dataset private-sensitive content identification should employ a two-pronged approach to improve the accuracy and flexibility of detection techniques. First, this calls for a concentrated effort to improve and enlarge the dataset that serves as the study's foundation. It is imperative to create a dataset that includes a greater variety of representative and diverse content samples from Norwegian social media. The inclusion of an expanded dataset would enhance the robustness of the model and provide a more precise representation of the language variety and contextual subtleties inherent in Norwegian online speech.

In addition to improving datasets, experimenting with different computational models has the potential to yield important advances in the discipline. Examining a wider range of models, such as sophisticated deep learning frameworks, transformer-based architectures, and advanced machine learning approaches, among others, may provide more effective and efficient methods for content detection. This investigation aims to investigate how new models might be incorporated or modified to enhance current approaches, rather than just evaluating their effectiveness through a rigorous and iterative process. Developing a thorough awareness of the possibilities of different computational techniques within the complex Norwegian social media environment is the ultimate goal of this kind of work, which will aid in the creation of models that are highly accurate, broadly applicable, and useful in real-life situations. Future study in these targeted areas can increase the identification of private-sensitive information and enhance the consideration of privacy concerns in the rapidly changing social media ecosystem.

In conclusion, continuous refinement of the dataset, exploration of diverse models, and domain-specific adaptations represent promising directions for future research in the field of private-sensitive content detection in Norwegian social media. These efforts aim to improve the models' accuracy, generalizability, and applicability to real-world scenarios.

ACKNOWLEDGEMENT

This paper is based on the master thesis from Norwegian University of Science and Technology titled "Detecting Private-Sensitive Content in Norwegian Social Media" by Haldis Borgen and Oline Zachariassen, 2023 [11].

REFERENCES

[1] T. Dowerah Baruah, "Effectiveness of Social Media as a Tool of Communication and Its Potential for Technology Enabled connections: a micro-level Study," International Journal of Scientific and Research Publications, vol. 2, no. 5, May 2012, Available: https://www.ijsrp.org/research_paper_may2012/ijsrp-may-2012-24.pdf

[2] "EUR-Lex - 32016R0679 - EN - EUR-Lex," eur-lex.europa.eu. https://data.europa.eu/eli/reg/2016/679/oj (accessed Aug. 08, 2023).

[3] D. Correa, L. Silva, M. Mondal, F. Benevenuto, and K. Gummadi, "The Many Shades of Anonymity: Characterizing Anonymous Social Media Content," Proceedings of the International AAAI Conference on Web and Social Media, vol. 9, no. 1, pp. 71–80, Aug. 2021, doi: https://doi.org/10.1609/icwsm.v9i1.14635.

[4] M. Petrolini, S. Cagnoni, and M. Mordonini, "Automatic Detection of Sensitive Data Using Transformer- Based Classifiers," Future Internet, vol. 14, no. 8, p. 228, Jul. 2022, doi: https://doi.org/10.3390/fi14080228.

[5] M. Sleeper et al., "I Read My Twitter the Next Morning and was Astonished" A Conversational Perspective on Twitter Regrets. 2022.

[6] P. Murmann and Farzaneh Karegar, "From Design Requirements to Effective Privacy Notifications: Empowering Users of Online Services to Make Informed Decisions," vol. 37, no. 19, pp. 1823–1848, Jun. 2021, doi: https://doi.org/10.1080/10447318.2021.1913859.

[7] M. A. Svanes and T. S. Gunstad, "Detecting and Grading Hateful Messages in the Norwegian Language," ntnuopen.ntnu.no, 2020. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2777836 (accessed Aug. 10, 2023).

[8] L. Bioglio and R. G. Pensa, "Analysis and classification of privacy-sensitive content in social media posts," EPJ Data Science, vol. 11, no. 1, Mar. 2022, doi: https://doi.org/10.1140/epjds/s13688-022-00324-y.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1912.07076, Oct. 2018.

[10] Y. Liu et al., "Roberta: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[11] H. Borgen and O. Zachariassen, "Detecting Private-Sensitive Content in Norwegian Social Media," ntnuopen.ntnu.no, 2023. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/3100364 (accessed Mar. 19, 2024).