

Natural Language Processing and Deep Learning Based Techniques for Evaluation of Companies' Privacy Policies

Saka John[1(✉)], Binyamin Adeniyi Ajayi[1], and Samaila Musa Marafa[2]

[1] Department of Computer Science, Nasarawa State University Keffi, Keffi, Nigeria
1johnsaka78@yahoo.com

[2] Department of Computer Science, Federal University Gusau, Gusau, Nigeria

Abstract. Companies' websites are vulnerable to privacy attacks that can compromise the confidentiality of data which, particularly in sensitive use cases like personal data, financial transaction details, medical diagnosis, could be detrimental and unethical. The noncompliance of companies with privacy policies requirements as stipulated by the various Data Protection Regulations has raised lot of concerns for users and other practitioners. To address this issue, previous research developed a model using conventional algorithms such as Neural Network (NN), Logistic Regression (LR) and Support Vector Machine (SVM) to evaluate the levels of compliance of companies to general data protection regulations. However, the research performance shows to be unsatisfactory as the model's performance across the selected core requirements of the legislation attained F1-score of between 0.52–0.71. This paper improved this model's performance by using Natural Language Processing (NLP) and Deep Learning (DL) techniques. This was done by evaluating the same dataset used by the previous researcher to train the proposed model. The overall results show that LSTM outperform both GRU and CNN models in terms of F1-score and accuracy. This research paper is to assist the Supervisory Authority and other practitioners to better determine the state of companies' privacy policies compliance with the relevant data protection regulations.

Keywords: Privacy policies · Natural language processing · Deep Learning · Manual Annotation · Machine Learning · Artificial Intelligent · Crowdsourcing

1 Introduction

Companies' privacy policies are legal document that are generally used to communicate how the personal information or data collected are used or managed. Accepting it means that users have agreed to release their personal information and data to the company as contained in the conditions specified by the company policy (Costante et al. 2012). Today, businesses around the world generally depend on data and other related information that are derived from people's data. Nowadays, data is the world's most important and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022 Workshops, LNCS 13377, pp. 15–32, 2022.
https://doi.org/10.1007/978-3-031-10536-4_2

valuable resources, it is no longer oil (Labadie and Legner 2019). Thus, data is of more value than many other assets.

However, due to the widespread breaches and leaks of data and then information, many users are now concerned about its privacy. As a result, they would like to know the privacy control that can be exercised by them in privacy policy (Sathyendra et al. 2017). To comply with provisions of data protection regulations, companies are now being forced to appraise and simplify their privacy policies to make their users or customers aware of how their data are collected, processed and used (Sánchez et al. 2021). Thus, data privacy has become the most significant issues in companies across the globe. There is need for transparency in how companies obtain consent, abide with their policies and effectively manage all data collected from the users. This is vital in building a strong trust and accountability with all customers and other partners who expect data privacy (Chika and Tochukwu 2020).

Many business owners and companies have taken steps to protect their user's data by drafting a concise and clear company's privacy policy that gives user a better understanding of how their data is being processed or used. This is expected to instill faith, trust and protects companies from a number of liability issues. Most countries have adopted or are in the course of adopting an all-inclusive data protection regulation similar to the General Data Protection Regulation (GDPR). In January 2019, Nigeria released a revised legislation called the Nigeria Data Protection Regulation (NDPR), which resemble the current GDPR. The NDPR was designed to completely change the corporate and other personal transactions within the country in order to strengthen data privacy and protection policies.

Table 1. GDPR Requirements (Muller et al. 2019)

No	Category	Required companies privacy policy content
1	DPO	Contact details for the data protection officer or Equivalent
2	Purpose	Disclosure of the purpose for which personal data is or is not used for
3	Acquired Data Disclosure	that personal data is or is not collected, and/or which data is collected
4	Data sharing	Disclosure if 3rd parties can or cannot access a user's personal data
5	Rights	Disclosure of the user's right to rectify or erase personal data

This study evaluates companies' privacy policies using NLP and DL based techniques so as to determine the state of various companies' privacy policies compliance with the data protection and privacy legislation. A study that has developed a similar model in the past is (Muller, Kowatsch, Debus, Mirdita & Böttinger, 2019). The authors created a dataset labelled in accordance to the core GDPR requirements to support practitioners in analysis and detection of policy non-compliance with the provision of GDPR as shown in Table 1.

However, their model's F1-scores performance achieved between 0.52–0.71 across the five core requirements of the data protection regulations. This study aims to improve this model performance by using NLP and DL as well as a more proficient word embedding technique to achieve a better result.

2 Related Works

Most of the existing literature in the area of data protection regulations and privacy policies tackles different issues of compliance, readability and clarity of the privacy policies. Most of the researchers employed the use of NLP, DL, Manual Annotation, Machine Learning, Artificial Intelligent, Crowdsourcing and Automatic Classification techniques.

A study, Connor (2004) analyzed companies' privacy statement and found that in most cases, they failed to comply with globally known standard. The study designed an intelligent software to carry out an in-depth structural analysis of privacy policies documents using content analysis techniques on a small data set. Also, to ensure the privacy of children's data Micheti et al. (2010) designed a privacy policies re-written framework and employed comparative analysis using statistical methods. Their framework encouraged teenagers to use personal data and information on company website.

Similarly, Costante et al. (2012) developed a solution to enable users make an informed decision while surfing the internet. The study used shallow machine learning approach to build text classifier. This classifier was tested by simulating a user browsing session and obtained an accuracy of 99.3% and 97.9% precision, but the recall was relatively. In Sadeh et al. (2013), the researcher showed that many users hardly read these policies and those who sometimes does most often struggle to understand its content. The study developed an algorithm that interpret companies' privacy policies using traditional ML and NLP. This approach only allows user to gain a deeper understanding of the privacy policies and to drive the privacy decision making processes.

To provide better understanding of the privacy policies documents, Zimmeck and Bellovin (2014) designed Privee - a software architecture for analyzing essential policy. It combines crowd sourcing with rule and machine learning classifiers to classify privacy policies that are not already rated in the crowd sourcing repository. The classifiers achieved an overall F1-score of 90%. However, this was done before the advent of GDPR.

In Liu et al. (2016), the author developed a DNN for vector representation of textual content in the companies' privacy policies documents. The study designed and analyzed vector representations for each of the word or token in the privacy policies documents. This was fed into the Long Short Term Memory Visualization tools (LSTMV) so as to test it capability to learn semantically and syntactically vague terms.

Due to the fact that the companies' privacy policy is usually very long, (Sathyendra et al. 2017) presented another way to detect text of choices in privacy policy documents. The study used corpus that consists of 115 website privacy policies and annotations created by law students. The authors achieved a F1-score of 0.735 for their classification model. This technique enabled the creation of systems to help internet users to learn about their choices, thereby improving data privacy.

Goltz and Mayo (2018) investigated the issue of non-compliance and proposed a framework to enhance regulatory compliance using Machine Learning (ML) to identify compliance clauses in the regulations. The study obtained an average accuracy of 90% for all approaches. Similarly, an automated framework for Privacy Policy Analysis Polisis, was introduced by Harkous et al. (2018). The study used Convolutional Neural Networks (CNNs) internally within all the classifiers to analyze and examine the depth and breadth of companies' privacy policies. This was done to summarize what Personally Identifiable Information, PII, that the companies' privacy policies documents claim to be requesting for, collecting from their users and sharing with other partners.

Zaeem et al. (2018) designed a browser add-on called Privacy Check that gives graphical and at-a-glance detailed summary of companies' privacy policies textual documents. Privacy Check used data mining tool to analyze textual content of the privacy policies documents and answers the ten (10) basic questions in respect of user's data privacy and data security as well as the information that are being collected from the users and how they are being used.

Another study, Andow et al. (2019) developed Policy Lint with the aid of NLP tool to find potential contradictions in the privacy policies text of various company. The study analysis focused on policies of Android apps. Privacy Guide was developed by Tesfay et al. (2018) to examine the issues of privacy policies readability and other complexities associated with the document. This guide was designed based on ML and NLP techniques. However, due to the problem of extracting companies' privacy policies online using web scrapers, Degeling et al. (2018) developed a system that can automatically scan websites for links to firm's privacy policies documents. The study used NLP tools to perform the analysis on the word-level and check if companies' privacy policies documents stated the phrases that are specifically mentioned in the data protection and privacy regulations.

In all the foregoing related studies, the most closely related work to the proposed study is Müller et al. (2019). The study introduced new labeled dataset of companies' privacy policies to study GDPR compliance. This dataset comprises of 250 companies' privacy policies text of different firms and over 18,390 sentences, labeled across the five (5) major classes of Acquired Data, Rights, DPO, Data Sharing and Purpose, were applied to a set of NLP and ML algorithms to analyze the dataset. The study achieved F1-scores of 0.52–0.71 across these five classes of data protection requirements. However, this result shows there is room to improve their work with respect to word embedding techniques and the classification algorithm. The subsequent sections highlight the procedure employed to obtained the set goals.

3 Methodology and Proposed Architecture

This section focusses on the research designs and techniques that were adopted in this study with the aim of achieving the research objectives. The goal of this study is to improve the performance of the model developed by Müller et al. (2019). Doing this, it developed a model that increase the F1-score and also address observed methodological gaps in their work.

The proposed model was experimented on the same dataset used by Müller et al. (2019). It is informative that the author used NLP tool and ML algorithms such as SVM,

LR and NN. However, the model's performance seems to be unsatisfactory as it only achieved a classification accuracies F1-scores between 0.52 to 0.71 across the five core requirements of data protection regulations. According to Franzoni and Kozak (2021), since performances of classifiers based on LR and SVM were not particularly remarkable, the possibility to improve their precision has been explored, applying different techniques of feature extraction and feature selection.

This research work relies mostly on the secondary data sourced from the first large-scale dataset created by (Müller et al. 2019) for GDPR compliance test. Using mathematical models or deep learning approaches, they result highly efficient on basic models (Baia et al. (2022). Hence, the proposed solution seeks to improve the model using NLP techniques, GloVe Word Embedding techniques and DL algorithms. Thus, DL algorithms availed include Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU) and Convolutional Neural Network (CNN) were used for classification of companies' privacy policies documents based on core requirements of data protection regulations.

3.1 Proposed Companies' Privacy Policies Analysis Framework

Companies' websites usually disclose new rules, products, information and technologies that are being implemented. Although, this information is generally available to the public, when extracted from the organization's website, it is usually unstructured (Kinne and Axenbeck 2018). Hence, there is need to refine this information with a more reliable methodology for collection or gathering and harmonizing companies' privacy policies text. This is to enable extraction of the core requirements of data protection and privacy regulations done with a view to check the companies' compliance with these legislations, which is the goal of this research work.

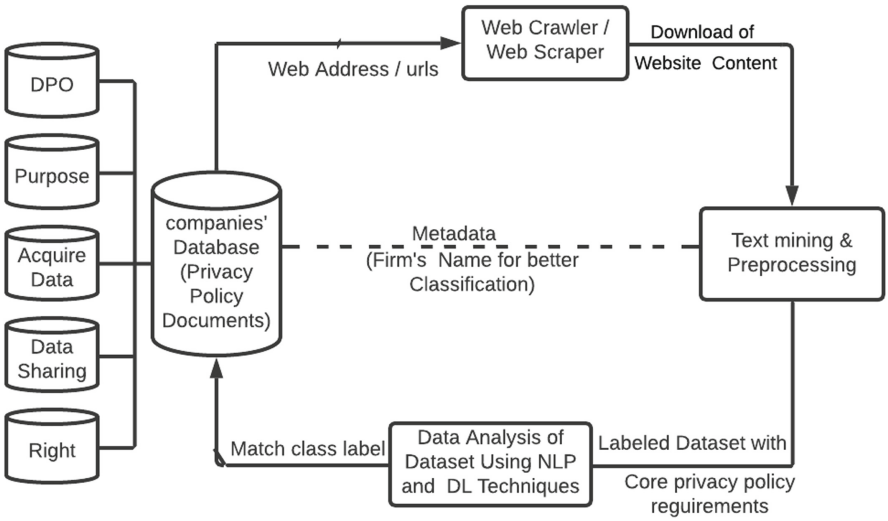

Fig. 1. Proposed Companies' Privacy Policies Analysis (CPPA) Framework. Adapted from Kinne and Axenbeck (2018)

Figure 1 outlined a methodology tagged Companies' Privacy Policies Analysis (CPPA) Framework adapted from Kinne and Axenbeck (2018). The framework is proposed for extracting and classifying the companies' privacy policies based on the identified core requirement of data protection and privacy regulation. The details components of this methodology were discussed in the next sub-section.

3.1.1 Web Crawler or Web Scraper

In the first step, web crawler or web scraper is to download the website content containing the privacy policies of various company. The main data and information that is extracted includes the privacy policy statement, company name and website addresses (URLs). However, since this study is to use the dataset created by Müller et al. (2019) there is no need to repeat this process.

3.1.2 Text Mining and Preprocessing

A text mining is also known as text data mining which is the process of deriving relevant information or information of interest for the classification of privacy policies documents so as to draw meaningful insights and patterns from the unstructured data. At this stage, the textual data that is extracted during the first stage is preprocess so as to clean the text document into a form that is predictable and analyzable for the task at hand. Data preprocessing organize a raw textual data for processing.

3.1.3 Data Analysis Using NLP and DL Techniques

At this point, relevant textual data was analyzed using a combination of NLP tools to extract needed data to match the class label. The top keywords were extracted using NLP tools. Thereafter, the DL techniques was applied to the labeled dataset to classify it according to the core requirements of data protection and privacy regulations. Based on this, data protection regulations evaluation is perform using DL algorithms.

3.1.4 Companies' Database

The companies' database contains the labelled dataset which represent the basic requirements of data protection and privacy regulations. This dataset was used for predictive classification purposes, or store in the companies' database. Furthermore, the dataset was analyzed to evaluate or examine the state of the organization's private policies with the data protection legislations.

3.1.5 Metadata

In this section, additional information about data is automatically mined and extracted from other relevant metadata of the companies' website. This information can be used to support the analysis, if necessary, to better enhance the preprocessing and data classification of the text data. In the final step, a model is trained in order to evaluate the state of companies' privacy policies using the core requirements of data protection and privacy regulations described in Table 1.

3.2 System Architecture

The proposed approach was developed using NLP, DL and sentence embedding pre-training techniques for the analysis and evaluation of companies' privacy policies documents. A deep neural network was trained using textual data which were extracted from websites of organizations for which established privacy policies are obtainable. These policies were used to create the training data set of labelled textual data.

According to Kinne and Axenbeck (2018) recent developments in the NLP gives DL approaches a more potentially and make it most promising when it comes to gathering data and information about processes and activity based on textual document content. In this study, the privacy policies text classification was carried out using NLP techniques and DL algorithms. Figure 2 illustrates the schema of the proposed model system architecture adapted from Kinne and Axenbeck (2018).

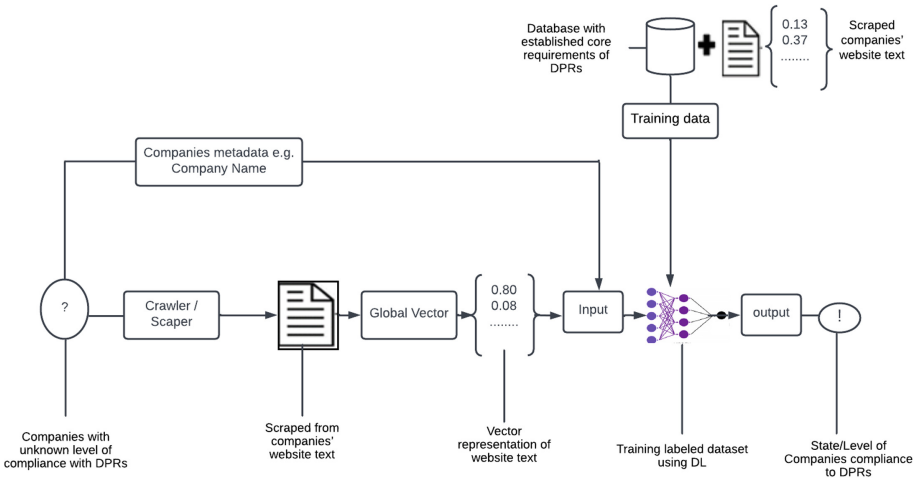

Fig. 2. Proposed Approach for the Evaluation of Companies Privacy Policies using NLP and DL Technique, Adapted from Kinne and Axenbeck (2018)

In Fig. 2, the first step is to extract the dataset from the companies' web data. The dataset contains information about the organization's privacy policy, url, name, location and address of the company. This dataset was used to train the model that checks if the companies' privacy policies contain core requirements of the data protection regulations. However, for the purpose of this study, a secondary dataset that has been extracted from companies with unknown level of data privacy compliance created by Müller et al. (2019) was adopted.

The second step deal with the analysis of the data set mined by the previous study from the companies' privacy policies. Likewise, a number of NLP tools was employed along with word embeddings techniques such as GloVe to map textual data to their respective vector representation. Each instance was mapped into a deep semantic vector representation and used as input to train the proposed model.

Finally, a DL algorithm and training data sets were used to build, develop and train the proposed model. Meanwhile, the information extracted from unseen companies' privacy policies (test data) were used to test the trained model. The model was then analyzed and evaluated to check if it contains the core requirements stipulated by DPR.

3.3 Programming Language

For the implementation of this study, the programming language of choice is Python version 3.6. Other software requirements were TensorFlow and Keras DL frameworks. This software was executed on Window Operating System environment deployed on HP Laptop, Core (TM)i3-5005U CPU with 4 GB of RAM and 2 GHz Processor Speed. All the required libraries were installed on the python note-book and used for the data analysis and models development.

4 System Implementation and Experimental Evaluation

Having reviewed the existing model proposed by Muller el at. (2019), the next thing is to implement the system using deep learning algorithms and GloVe Word Embeding techniques.

4.1 Splitting of Dataset into Training and Test Set

The dataset created by the pervious researchers which was downloaded from their http://git.aisec.fraunhofer.de/projects/GDPRCOM/repos/on-gdprcompliance was used to train the proposed model. After obtaining a clean data set, the data was divided into training and testing data sets. 70% of the whole data was apportioned for the training while 30% was used for testing of the model. The number of instances for the training and test set are shown in Table 2.

Table 2. Data split

Dataset	Number of sentences
Training	12877
Test	5520
Total	**18397**

4.2 Word Tokenization

Tokenization involves converting text into sequence of integers or numbers. This was applied in this study because statistical techniques such as machine learning can only deal with numbers. Therefore, it is important to convert the textual data into numbers as required. This study applied tokenizer from keras for the tokenization.

4.3 Word Embeddings

A word embeddings technique is a class of approach for the representation of word and document using a vector representation. GloVe was used in this study to obtain a vector representation for words in the textual data set. A word embeddings dimensions 100 vector was used. This means that each of the word was mapped to a 100-dimensioned float numbers. This mapping was adjusted during the training of the model.

4.4 Classification

To resolve the issue of imbalance, this study applied Synthetic Minority Oversampling Technique SMOTE (Bowyer et al. 2011). After performing SMOTE operation on the companies' privacy policies dataset, the training data and its corresponding labels were fed into the classification model. LSTM, CNN and GRU algorithms were used for the classification tasks. This was implemented on Keras framework built on tensorflow (backend) used to design framework. The LSTM, CNN and GRU units takes it parameters as a number of units which is the vector dimension of the output of each network blocks. The next subsection shows details of the learning process.

4.5 Experiments: DL Learning Curves

This subsection discusses the outcome of CNN, GRU and LSTM learning process during training. The plots below show the learning curves (accuracy and loss plot on training and validation sets) for both SMOTE and Non-SMOTE GDPR datasets. This was carried out on each category or parameter. Both accuracy and loss were plotted against 5 epochs. Figure 3 depicts the acquired data class using CNN algorithm.

Figure 3 shows the acquired data class by implementing the CNN algorithm.

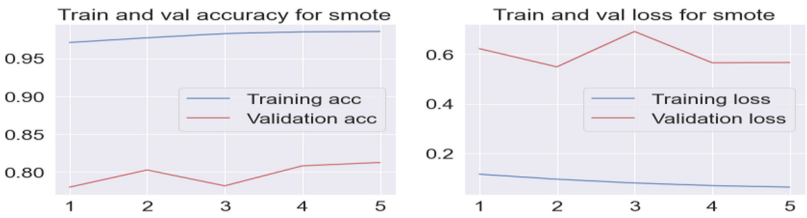

Fig. 3. Train and validation accuracy and loss of SMOTE using CNN

Figure 4 shows the acquired data class by implementing the GRU algorithm.

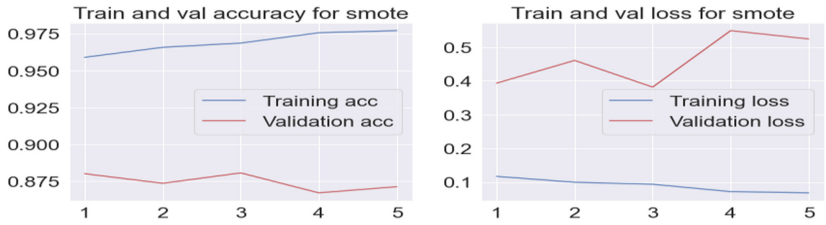

Fig. 4. Train and validation accuracy and loss of SMOTE using GRU

Figure 5 depicts the acquired data class by implementing the LSTM algorithm.

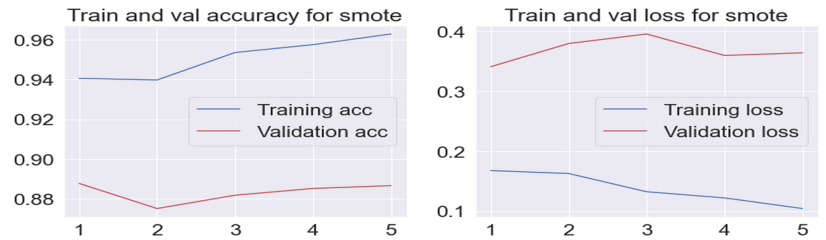

Fig. 5. Training for validation accuracy and loss of SMOTE using LSTM

Figure 6 depicts the DPO class using CNN algorithm.

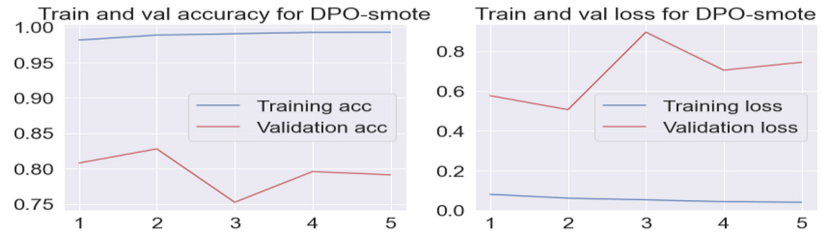

Fig. 6. Training for validation accuracy and loss of SMOTE using CNN

Figure 7 depicts the DPO class using the GRU algorithm.

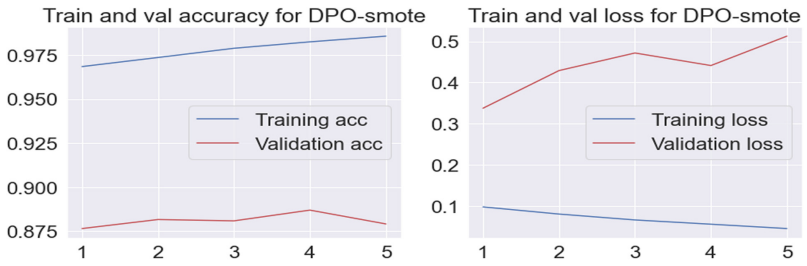

Fig. 7. Training for validation accuracy and loss of SMOTE using GRU

Figure 8 depicts the DPO class using the using LSTM algorithm.

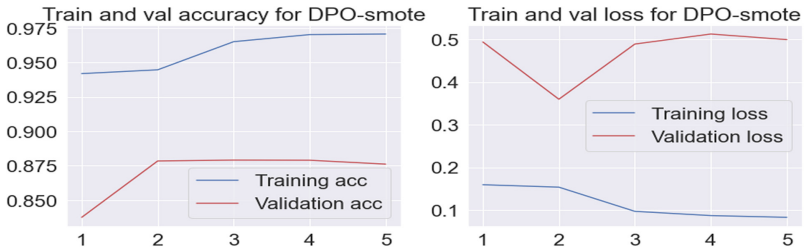

Fig. 8. Training for validation accuracy and loss of SMOTE using LSTM

Figure 9 is the Purpose class using the CNN algorithm.

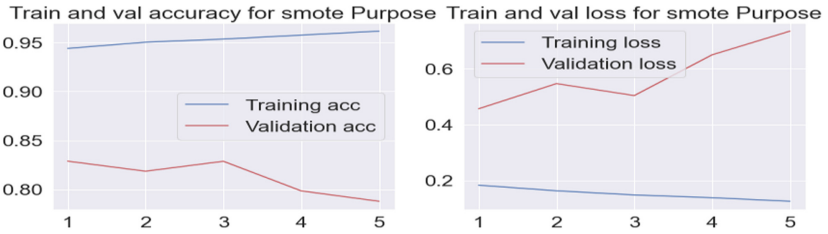

Fig. 9. Training for validation accuracy and loss of SMOTE using CNN

Figure 10 depicts the Purpose class using GRU algorithm.

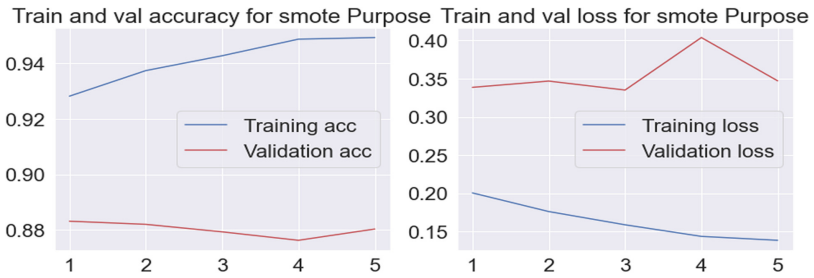

Fig. 10. Training for validation accuracy and loss of SMOTE using GRU

Figure 11 depicts the Purpose class using LSTM algorithm.

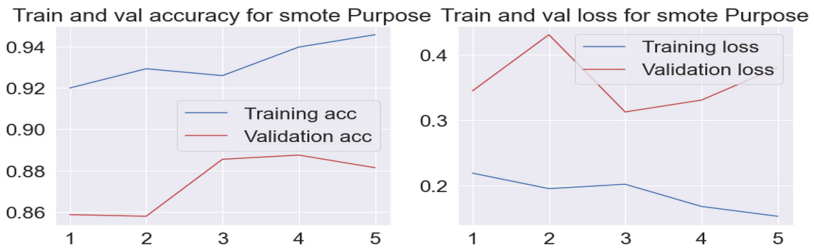

Fig. 11. Training for validation accuracy and loss of SMOTE using LSTM

Figure 12 depicts the Rights Class using CNN algorithm.

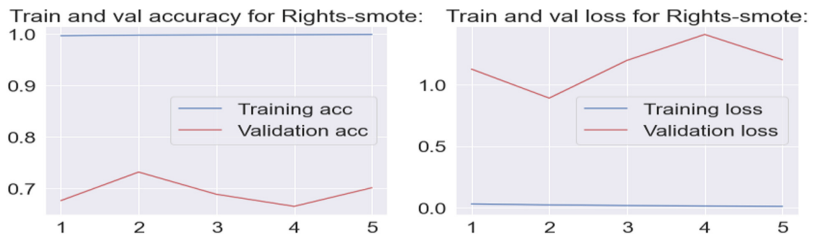

Fig. 12. Training for validation accuracy and loss of SMOTE using CNN

Figure 13 depicts the Rights class using GRU algorithm.

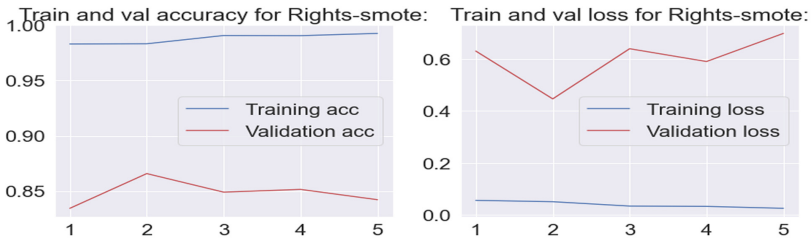

Fig. 13. Training for validation accuracy and loss of SMOTE using GRU

Figure 14 depicts the Rights class using the LSTM algorithm.

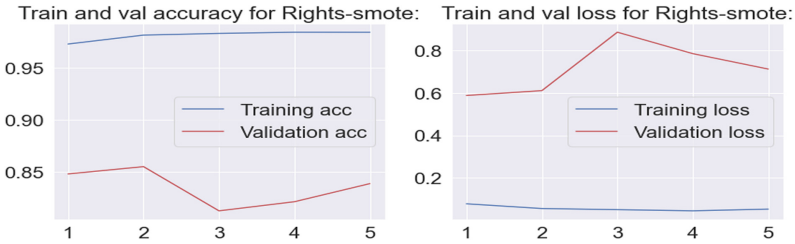

Fig. 14. Training for validation accuracy and loss of SMOTE using GRU

Figure 15 depicts the Data Sharing (DS) class using the CNN algorithm.

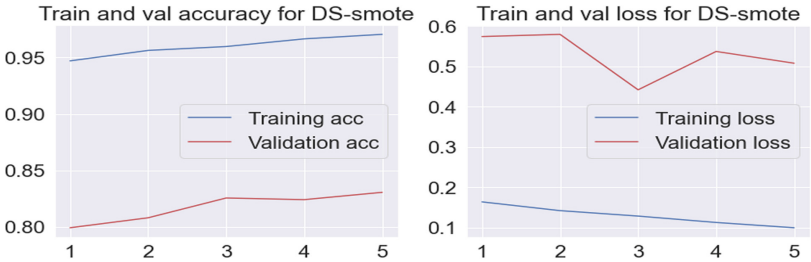

Fig. 15. Training for validation accuracy and loss of SMOTE using CNN

Figure 16 depicts the Data Sharing (DS) class using the GRU algorithm.

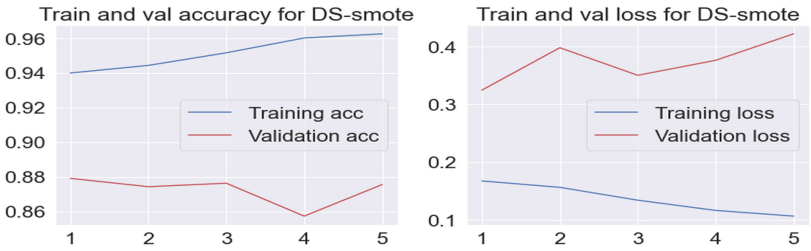

Fig. 16. Training for validation accuracy and loss of SMOTE Using GRU

Figure 17 depicts the Data Sharing (DS) class using the LSTM algorithm.



Fig. 17. Training for validation accuracy and loss of SMOTE using LSTM

4.6 Experimental Results and Discussion

This section shows the evaluation and result of the developed model on the test data set. The performance metric applied includes the F1-score, accuracy, precision and recall whose parameters were obtained from confusion matrix computation. According to Ramaiah et al. (2021), these are commonly used metrics to measure the performances of the ML models. Confusion matrix is a table that is often used to describe the performance of a class. It shows the ways in which the model is confused when it makes predictions and depicts insights, not only into the errors being made by the classifier but more importantly the types being made.

Tables 3, 4 and 5, depicts that the CNN, GRU and LSTM performances on SMOTE companies' privacy dataset was very low despite the high accuracy score. The poor recall, precision and F1-score show that the model favoured more of the majority class set than the minority class. However, the model performed better on balanced data using SMOTE with a higher precision, recall, and F1-score. Table 3 shows CNN performances on SMOTE companies' privacy dataset.

Table 3. Performance evaluation of the model on test data using CNN

Categories	CNN			
	Accuracy	Precision	Recall	F1-score
Acquire Data	0.80	0.81	0.66	0.73
DPO	0.78	0.79	0.65	0.71
Purpose	0.75	0.65	0.81	0.73
Data sharing	0.79	0.74	0.77	0.75
Right	0.74	0.90	0.42	0.57

Table 4 shows GRU performances on SMOTE companies' privacy dataset.

Table 4. Performance evaluation of the model on test data using GRU

Categories	GRU			
	Accuracy	Precision	Recall	F1-score
Acquire data	0.87	0.98	0.70	0.82
DPO	0.88	0.97	0.73	0.83
Purpose	0.88	0.94	0.76	0.84
Data Sharing	0.88	0.93	0.76	0.83
Right	0.84	0.97	0.63	0.77

Table 5 shows LSTM performances on SMOTE companies' privacy dataset.

Table 5. Performance evaluation of the model on test data using LSTM

Categories	LSTM			
	Accuracy	Precision	Recall	F1-score
Acquire data	0.86	0.84	0.81	0.83
DPO	0.89	0.93	0.78	0.85
Purpose	0.87	0.99	0.68	0.81
Data sharing	0.88	0.92	0.78	0.84
Right	0.84	0.96	0.63	0.76

The overall results show that LSTM outperform both GRU and CNN models. LSTM achieves F1-scores of 0.76–0.85 and accuracy of 0.84–0.89. Likewise, GRU's F1-scores stood at 0.77–0.84 and accuracy of 0.84–0.88. CNN achieves a F1-scores of 0.57–0.75

and accuracy of 0.74–0.80 which is an improvement over the previous study. Figure 18 shows the F1-scores of the CNN, LSTM and GRU algorithms.

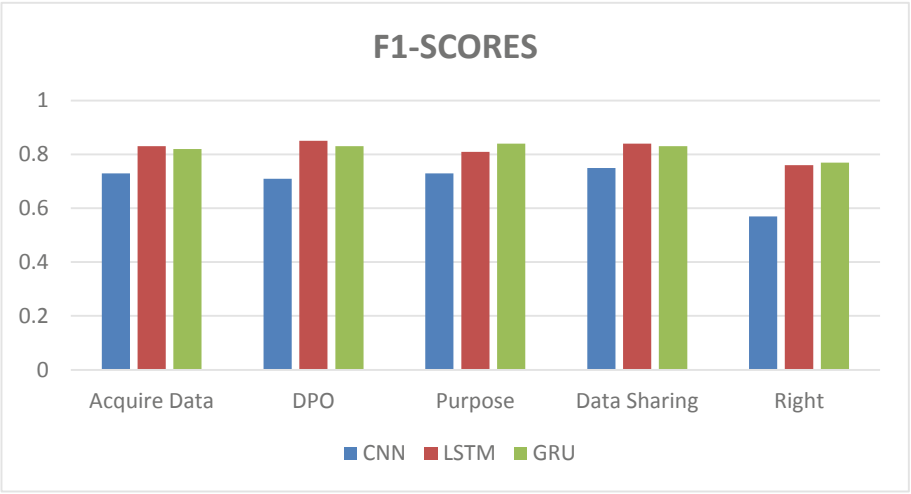

Fig. 18. Comparing F1-Scores of CNN, LSTM and GRU

Figure 19 shows the model accuracy of the CNN, LSTM and GRU algorithms.

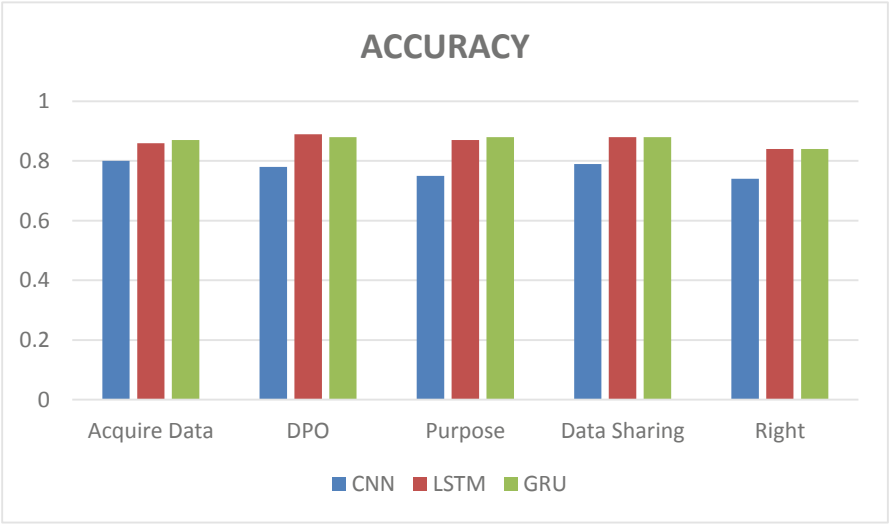

Fig. 19. Comparing accuracy of CNN, LSTM and GRU

5 Conclusion

This research presents a solution to better evaluate companies' privacy policies, using Natural Language Processing and Deep Learning techniques. This study shows that, given the significance and complexity of companies' privacy policies, being able to automatically evaluate their completeness would give significant benefits to the end-users and other key practitioners. This study used CNN, LSTM and GRU algorithms over a corpus of companies' privacy policies dataset, to prove the viability of the proposed model. The overall results across the five selected core requirement of companies' privacy policies shows that LSTM outperform both GRU and CNN models. The LSTM achieves an F1-scores of 0.76–0.85 and accuracy of 0.84–0.89. Likewise, GRU achieves an F1-scores of 0.77–0.84 and accuracy of 0.84–0.88 while CNN achieves an F1-scores of 0.57–0.75 and accuracy of 0.74–0.80 which is an improvement over the previous model using the same dataset.

The outcome of the proposed model shows an accurate classification framework for the analysis of companies' privacy policies. It is expected that this framework will give an effective result whenever Supervisory Regulators and other practitioners want to evaluate the state of compliance of company privacy policy with the core requirement of the data protection and privacy regulations. Such policies and regulations are as enacted by various countries across the globe.

The model testing and performance evaluation results show an excellent improvement than the existing model and classification systems. Therefore, the CPPA model implementation is recommended to overcome the observed methodological gaps and inadequacies inherent in previous model and systems of operation.

References

Andow, B., et al.: Policylint: investigating internal privacy policy contradictions on google play. In: 28th USENIX Security Symposium (USENIX Security 19), pp. 585–602 (2019)

Baia, A.E., Biondi, G., Franzoni, V., Milani, A., Poggioni, V.: Lie to me: shield your emotions from prying software. Sensors **22**(3), 967 (2022)

Bowyer, K.W., Chawla, N.V., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority oversampling technique. CoRR abs/1106.1813 (2011). http://arxiv.org/abs/1106.1813

Costante, E., Sun, Y., Petković, M., Den Hartog, J.: A Machine Learning Solution to Assess Privacy Policy Completeness: (short paper). In: Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society, pp. 91–96, October 2012

Chika, D.M., Tochukwu, E.S.: An Analysis of Data Protection and Compliance in Nigeria (2020). https://www.rsisinternational.org/journals/ijriss/DigitalLibrary/volume-4-issue-5/377-382.pdf

Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., Holz, T.: We value your privacy... now take some cookies: Measuring the GDPR's impact on web privacy. arXiv preprint arXiv: 1808.05096 (2018)

Franzoni, V., Kozak, Y.: Yeasts automated classification with extremely randomized forests. In International Conference on Computational Science and Its Applications, pp. 436–447. Springer, Cham, September 2021

Goltz, N., Mayo, M.: Enhancing regulatory compliance by using artificial intelligence text mining to identify penalty clauses in legislation. RAIL **1**, 175 (2018)

Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K. G., Aberer, K.: Polisis: Automated analysis and presentation of privacy policies using deep learning. In: 27th USENIX Security Symposium (USENIX Security 18), pp. 531–548 (2018)

Kinne, J., Axenbeck, J.: Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study for Germany. In: ZEW-Centre for European Economic Research Discussion Paper, (18–033) (2018)

Micheti, A., Burkell, J., Steeves, V.: Fixing broken doors: strategies for drafting privacy policies young people can understand. Bull. Sci. Technol. Soc. **30**(2), 130–143 (2010)

Muller, N. M., Kowatsch, D., Debus, P., Mirdita, D., Böttinger, K. (2019, September). On GDPR Compliance of Companies' Privacy Policies. In: International Conference on Text, Speech, and Dialogue, pp. 151–159. Springer, Cham (2019)

Labadie, C., Legner, C.: Understanding data protection regulations from a data management perspective: a capability-based approach to EU-GDPR. In: Proceedings of the 14th International Conference on Wirtschaftsinformatik, February 2019

Liu, F., Fella, N. L., Liao, K.: Modeling Language Vagueness in Privacy Policies Using Deep Neural Networks. In: 2016 AAAI Fall Symposium Series, September 2016

O'Connor, P.: Privacy and the online travel customer: an analysis of privacy policy content, use and compliance by online travel agencies. In: ENTER, pp. 401–412, January 2004

Ramaiah, M., Chandrasekaran, V., Ravi, V., Kumar, N.: An intrusion detection system using optimized deep neural network architecture. Trans. Emerging Telecommun. Technol. **32**(4), e4221 (2021)

Sadeh, N., et al.: The usable privacy policy project: Combining crowdsourcing. Machine Learning and Natural Language Processing to Semi-Automatically Answer Those Privacy Questions Users Care About. Carnegie Mellon University Technical Report CMU-ISR-13–119, 1–24 (2013)

Sathyendra, K.M., Wilson, S., Schaub, F., Zimmeck, S., Sadeh, N.: Identifying the provision of choices in privacy policy text. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2774–2779, September 2017

Sánchez, D., Viejo, A., Batet, M.: Automatic assessment of privacy policies under the GDPR. Appl. Sci. **11**(4), 1762 (2021)

Tesfay, W.B., Hofmann, P., Nakamura, T., Kiyomoto, S., Serna, J.: PrivacyGuide: Towards an Implementation of the EU GDPR on Internet Privacy Policy Evaluation. In: Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics, pp. 15–21, March 2018

Zaeem, R.N., German, R.L., Barber, K.S.: Privacycheck: automatic summarization of privacy policies using data mining. ACM Trans. Internet Technol. (TOIT) **18**(4), 1–18 (2018)

Zimmeck, S., Bellovin, S.M.: Privee: an architecture for automatically analyzing web privacy policies. In 23rd Security Symposium (USENIX Security 14), pp. 1–16 (2014)