Deep Learning Approach for Classifying the Aggressive Comments on Social Media: Machine Translated Data Vs Real Life Data

shapna akter (jannatul.shapna99@gmail.com)
North South University
Nova Ahmed
North South University

Research Article

Keywords: Cyber-agression, Augmentation, Data preprocessing, AutoEncoder, LSTM, BiLSTM, Word2vec, BERT, GPT-2

Posted Date: April 29th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1519060/v1

Deep Learning Approach for Classifying the Aggressive Comments on Social Media: Machine Translated Data Vs Real Life Data

Mst. Shapna Akter[1] and Nova Ahmed[1]

[1]Department of Electrical Computer Engineering, North South University, Bashundhara, Dhaka, 1229, Bangladesh.
{shapna.akter, nova.ahmed}@northsouth.edu
[1]Corresponding author's email: shapna.akter@northsouth.edu

Abstract

Aggressive comments on social media negatively impact human life. Such offensive contents are responsible for depression and suicidal-related activities. Since online social networking is increasing day by day, the hate content is also increasing. Several investigations have been done on the domain of cyberbullying, cyberaggression, hate speech, etc. The majority of the inquiry has been done in the English language. Some languages (Hindi and Bangla) still lack proper investigations due to the lack of a dataset. This paper particularly worked on the Hindi, Bangla, and English datasets to detect aggressive comments and have shown a novel way of generating machine-translated data to resolve data unavailability issues. A fully machine-translated English dataset has been analyzed with the models such as Long Short term memory model (LSTM), Bidirectional Long-short term memory model (BiLSTM), LSTM-Autoencoder, word2vec, Bidirectional Encoder Representations from Transformers (BERT), and generative pre-trained transformer (GPT-2) to make an observation on how the models perform on a machine-translated noisy dataset. We have compared the performance of using the noisy data with two more datasets, such as raw data, which does not contain any noises, and semi-noisy data, which contains a certain amount of noisy data. We have classified both the raw and semi-noisy data using the aforementioned models. To evaluate the performance of the models, we have used evaluation metrics such as F1-score, accuracy, precision, and recall. We have achieved the highest accuracy on raw data using the gpt2 model, semi-noisy data using the BERT model, and fully machine-translated data using the BERT model. Since many languages do not have proper data availability, our approach will help researchers create machine-translated datasets for several analysis purposes.

Keywords: Cyber-agression, Augmentation, Data preprocessing, AutoEncoder, LSTM, BiLSTM, Word2vec, BERT, GPT-2.

1 Introduction

The invention of the internet makes our life more accessible than before. The world is getting smaller day by day. We can communicate with others very easily from one corner of the world to another within a minute. The internet has brought us both positive and negative impacts on our life. Humans tend to use social media platforms because they can share their personal information, educational content, entertaining content, and

many more. The most popular social media platforms used in Asian regions and other countries are Facebook, Twitter, Instagram, and Youtube. Those platforms allow people to share their thoughts, information, and videos at no cost. The activities on virtual social media end up with negative impacts such as depression and suicide. Everyone is free to give their opinions. As a result, some people are getting aggressive towards the victims without worrying about their feelings. The commenter does not think about the intensity of negative comments, and thus they attack victims very harshly. Hence, the social media platform is now getting dangerous for the users and may lead some people to commit suicide. Many investigations have been done on cyberbullying and hate speech to find out the negative comments for taking some actions. Since the number of users increases daily, the proper investigation has become necessary to solve cyberbullying issues. However, countries like Bangladesh and India lack proper investigations due to the lack of data availability. None of the investigations has been done to resolved the data unavailability issues. Therefore, those countries are more vulnerable to cyberbullying and cyber aggression due to the lack of research and inappropriate action. We have focused on Hindi, Bangla, and English to detect aggressive comments on the social media platform. We have used the TRAC-2 dataset, which contains English, Bangla, and Hindi comments. Furthermore, we have tried to resolve the data unavailability issues by creating a fully machine-translated English dataset. Since, the dataset is very important to learn a model for detecting the aggressive comment, if one language lacks the dataset, it may become impossible to solve the aggressive comment issue for that language. Google Translate can translate data from one language to another but contains a lot of noises that may not be appropriate for training a deep learning model. In this paper, we have shown how the Deep learning models perform on machine-translated noisy datasets and compared the result with the raw and semi-noisy datasets, which have provided a clear observation of how the deep learning models perform on noise-free, semi-noisy, and fully-noisy datasets. The raw dataset contains a data imbalance issue. We have used a Machine translated augmentation process to avoid the imbalance problem and constructed the semi-noisy dataset. After constructing with all types of data, we have extracted features using the Bert embedding model. The extracted features have been fed to the deep learning models such as LSTM, BiLSTM, LSTM-Autoencoder, Word2vec, BERT, and gpt2 model for classifying the aggressive comments. We have checked the performance of the models using performance metrics such as F1-score, precision, recall, and accuracy. All of the evaluations have been done on the unseen dataset. We have got the highest result accuracy of 80 percent on English raw data, 73 percent on Bangla raw data using the gpt2 model. For the semi-noisy dataset, we have got highest accuracy of 75 percent on English, 71 percent on Bangla, and 0.68 percent on the Hindi dataset using the BERT model. For the fully machine-translated English dataset, we have achieved the highest accuracy of 78 percent using the BERT model.
Following our approach, suicidal activities may reduce to an extent by detecting the aggressive comments and taking action based on the prediction. Our approach can be useful for languages that lack data availability.

1.1 Related Work

Deep learning models such as Word2vec, LSTM, BiLSTM, BERT, XLM-Roberta, and FastText are popular models to deal with textual data. Some of the models can capture the true meaning of the sentence very well; some are good at less computation process.
R. kumar et al. showed an LSTM model for detecting aggressive comments on social media [1]. They used the combined data of both Trac-1 and Trac-2 workshops. They classified the Non-Aggressive (NAG), Covertly Aggressive, and Overtly Aggressive using SVM, LSTM, and deep neural networks. K. kumari and J. P. Singh proposed an LSTM model with FastText and One-hot embeddings. They found that the FastText embedding with LSTM performed better than the One-hot embeddings [2]. L. S. M. Altın et al. proposed a BiLSTM model for classifying the textual data [3]. J. Lilleberg et al. proposed a word2vec model for text classification as word2vec brings extra semantic meaning [4]. L. Wensen et al. also used the word2vec model

for short text classification [5]. First of all, they built the semantic relevant concepts sets of Wikipedia, and then they applied the Word2vec model to measure the semantic similarity between the concepts. Similarly, E. M. Alshari et al. proposed the word2vec for sentiment analysis [6]. T. Ranasinghe and M. Zampieri proposed a Crosslingual transformer called XML-R to classify the aggressive comments on social media [7]. They trained the transformer with the Trac-1 dataset, saved the weights, and applied the weights for the Trac-2 dataset. They approached the method for solving the low resource dataset. A Bert-based transformer is shown in F. Ramiandrisoa, and J. Mothe's paper, who used the model for classifying the text data [8]. They classified the aggressive comments using BERT-large, which is composed of 24 BERT Layers. S. K. Tawalbeh et al. also showed a fined tuned BERT model for classifying the text data [9]. H. Liu et al. proposed a Bert-based ensemble learning approach [10]. S. Tawalbeh et al. used the BERT transformer to compare their proposed XGB-USE model [11]. M.-A. Tanase et al. proposed several pre-trained language transformer models for classifying Spanish datasets [12].

2 The Dataset

The dataset used in this work is collected from Trac-2 (workshop on trolling, Aggression, and cyberbullying), which provides English, Bangla, and Hindi datasets [13]. The Shared task contains Sub-task A (Aggression Detection) and Sub-Task B (Misogyny Aggression detection). Sub-task A contains three classes: Non-Aggressive (NAG), Overtly Aggressive (OAG), Covertly-Aggressive (CAG). The indirect aggression comment is labeled as Covertly-Aggressive (CAG), the direct Aggression comment is labeled as Overtly Aggressive (OAG), and none of the aggression comments is labeled as Non-aggressive (NAG). The subTask-B contains two classes: GEN and NGENA. A comment which indicates a man, woman, or transgender is labeled as GEN, and a comment that does not indicate gender is labeled as NGEN. All three Datasets contain both train and test sets. In our project, we experimented with Sub-task A, since Sub-task A's features fully align with our targeted prediction, which is aggression detection. The statistics of the dataset provided by the organizations have been shown in Table-1 for sub-Task A.

TABLE-1 : Label distribution of dataset for subtask A

Set	NAG	OAG	CAG	Total
English Training	3375	453	435	4263
English Testing	836	117	113	1066
Hindi Training	2245	829	910	3984
Hindi Testing	578	211	208	997
Bangla Training	2078	898	850	3826
Bangla Testing	522	218	217	957

Some examples of the text data is shown in figure 1.

English:

'These types of people should live in Pakistan in Peace' NAG

'She is not a sane person...She wants to break india.' OAG

'bitches spoil a men life', CAG

Hindi:

'आपकी हिन्दी और अंग्रेजी दोनों भाषाओं पर शानदार पकड़ है।', NAG

'शब्द अच्छे हैं पर आप से नही हो पा रहा मुबारक ही सही लगते हैं', OAG

'बिल्कुल आरएसएस विचारधारा है भाई तेरी...गोडसे भगवान है तुम लोगो का.....
महिला ना\nहोती तो तू भी ना इस धरती पर होता....फूलन देवी को बैन करने वाले CAG
भी तो तूम जैसे\nलोग थे...शूद्र को बैन करने वाले भी तुम जैसे संघी थे.....
तुम सावरकर जैसे\nदलाल के चमचे से यहीं उम्मीद की जा सकती है,.....
..अपने चैनल का नाम मोदी भक्त रख\nले.....'

Bangla:

'সুখের সংসারে আগুন লাগাই মেয়েরাই লোভে।' NAG

'বেঈমান টাকে চুল ধরে টানতে টানতে আবার স্টেশনে নিয়ে আসা উচিত।' OAG

'এক কথায় ও একটা অসভ্য নোংরা খারাপ মহিলা।', CAG

Figure 1: Example of dataset with categories

3 Methodology

We have done a data augmentation process on the raw dataset provided by the organization. The Augmentation process is briefly described below :

3.1 Data Augmentation

To create semi-noisy data, we have added noises with the raw data until the dataset resolves imbalanced issues. The data we have used is highly imbalanced for sub-task A. The imbalanced data performed poorly for predicting the aggressive comments. The category NAG holds 50 percent of the total text data, and the category OAG and CAG hold the other 50 percent of total text data. To resolve the imbalanced data issue, we have Augmented the text data in such a way that all of the classes maintain almost the same amount of text data. We have adopted two methods for the Augmentation process:
1. Adding Noises
2. Data Translation

3.2 Adding Noise

Noises are added by replacing a word with synonyms or antonyms, adding random stop words, shuffling some words on raw text data. The process helps to increase the corpus size while keeping the context same as the raw dataset.

3.3 Data Translation

Data have been translated from one language to another, e.g. English to Bangla, using google translator. All the languages have the same sub-tasks and classes. So, we have translated the texts with all possible combinations until the dataset reaches a balanced position. We translated the texts for NAG, OAG, and CAG classes from Sub-Task A.

We have added texts from the noise augmentation process and texts from the translation augmentation process into the raw data so that the new corpus holds an almost equal number of text data for Sub-task A . The statistics of the dataset after adding the Augmented data with raw data shown in Table-1 for sub-Task A

TABLE-2 : Label distribution for Augmented + Raw dataset used for Subtask:A

Set	NAG	OAG	CAG	Total
English Training	3375	2251	2546	8172
English Testing	836	117	113	1066
Hindi Training	2245	3497	1810	7552
Hindi Testing	578	211	208	997
Bangla Training	2078	1959	1966	6003
Bangla Testing	522	218	217	957

3.4 Fully Machine Translated Data

Using the translation augmentation process, we have generated a complete machine-translated English dataset that is fully noisy. We have translated the Bangla and Hindi Sub-Task A texts into English for creating the new dataset.

TABLE-3 : Label distribution for fully translated English dataset for Subtask:A

Set	NAG	OAG	CAG	Total
English Training	4373	3096	2588	10057

Some examples of the Machine translated English data is shown in figure 1.

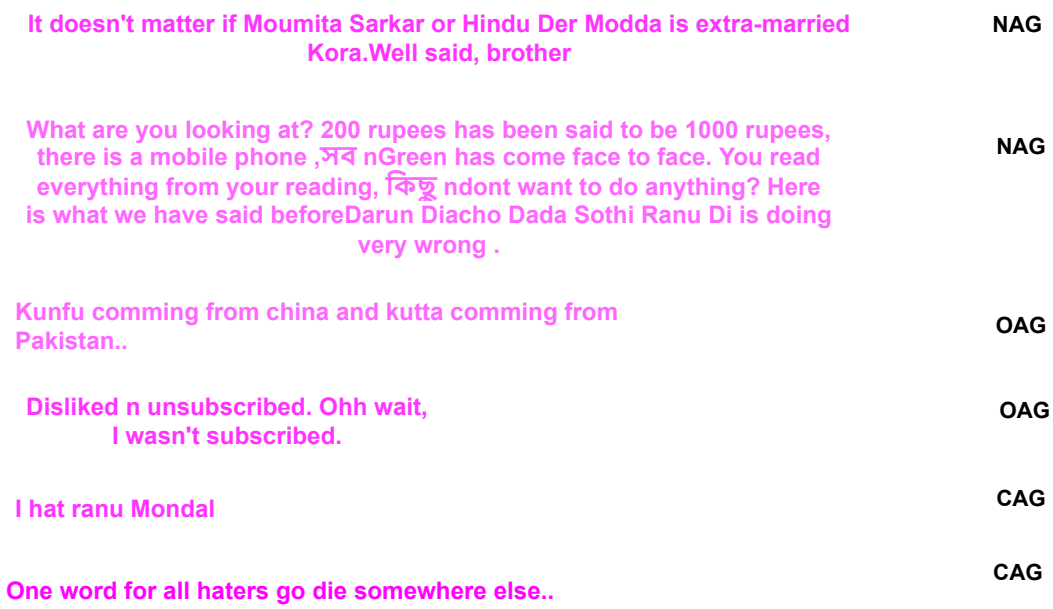

Figure 2: Example of Fully translated English dataset with categories

3.5 Input Representation

The raw text, semi-noisy, and fully noisy datasets are converted into a machine-understandable number representation. The computer can only understand the numbers; so, it is necessary to convert the text into numbers before feeding it into the models. We have used a BERT tokenizer for BERT models and TensorFlow.Keras tokenizer for Autoencoder, LSTM, and BiLSTM models.

3.6 Classification Models

The LSTM, BiLSTM, LSTM-autoencoder, Word2vec, and Bert models have been applied to the text data of trac-2 workshop. All of the models are individually applied on English, Bangla and, Hindi datasets . For training the models, the dataset has been split into two parts: training and validation. Finally, all of the trained models are evaluated on the test dataset.

3.6.1 LSTM

LSTM is a widely used artificial Recurrent neural network(RNN) model which deals with textual data. LSTM is the modified version of Simple RNN and can remember previous data points. The basic architecture of LSTM is developed by following the RNN model. However, LSTM is capable of learning long-term dependencies in the text dataset, which was one of the drawbacks of simple RNN [14]. We have used this model for classifying the aggressive comments.

3.6.2 BiLSTM

BiLSTM model is proposed by GRAVES[15]. The underlying concept of BiLSTM is the same as LSTM, except for the propagation from both sides. BiLSTM architecture learns from both the past to future direction. The backward propagation layer is basically the reverse layer of forwarding LSTM. This concept makes the architecture more stable since it works from both sides.

3.6.3 LSTM-Autoencoder

An LSTM Autoencoder is an autoencoder implemented for sequential data by following an Encoder-Decoder LSTM architecture. For a given sequential dataset, LSTM-Autoencoder was designed to read the input sequences, encode the sequence, decode the sequence, and reconstruct the sequence. The model's performance is estimated based on the ability to how correctly it can reconstruct the sequence. LSTM autoencoder can be used on video, text, audio, and time-series sequence data [16].

3.6.4 Word2vec

word2vec is a word embedding model which deals with textual data. Word Embedding is a type of word representation process which allows machine learning algorithms to understand words with similar meanings. It maps words into vectors of real numbers using neural network model and is capable of capturing a large number of precise syntactic and semantic word relationships. The neural network is two layered, which can detect synonymous words and suggest additional words for partial sentences once it is trained [17].

3.6.5 BERT

BERT stands for Bidirectional Encoder Representations from Transformers. It is a popular attention model for language modeling. Bert model looked at a text sequence either from left to right or combined left-to-right and right-to-left training, which ensures a deeper sense of language context and flow than single-direction language models. This paper has used two types of bert models: BERT base and BERT MultiLingual. Since Bert model is trained on English text data, leaving low-resource languages such as Bangla and Hindi behind. Whereas, Multilingual Bert model was trained on Wikipedia content with a shared vocabulary across all languages, which supports 104 languages. Bangla and Hindi dataset has been classified using the Multilingual Bert model [18].

3.6.6 GPT-2

Generative pretrained ransformer-2 (GPT-2) is one of the most standard states of the art generative modeling transformers. It has been trained on a large web text corpus. GPT2 is mostly used for the next sequence prediction, question answering, sequence classification, abstract, or text summarization. GPT 2 is known as a transformer decoder as it does not construct with lots of encoders; instead, it relies mainly on decoders as its main structural framework. GPT2 has many variances such as GPT-2 SMALL, GPT-2 MEDIUM, GPT-2 LARGE, GPT-2 EXTRA LARGE. We have used GPT-2 MEDIUM for classifying the aggression detection [19].

3.6.7 Evaluation metrics

Evaluating a model's performance is necessary since it gives an idea of how close the model's predicted outputs are to the corresponding expected outputs. The evaluation metrics are used to evaluate a model's performance. However, the evaluation metrics differ with the types of models. The types of models are classification and regression. Regression refers to the problem that involves predicting a numeric value. Classification refers to the problem that involves predicting a discrete value. The regression problem uses the error metric for evaluating the models . Unlike the regression problem model, The classification problem uses the accuracy metric for evaluation. Since Our motive is to classify the aggressive comments, we used Accuracy and F1 score as the main Evaluation metric. [20–23].

Precision : When the model predicts positive,it should be specify that how much the positive values are correct.Precision used when the false positive are high.In aggressive detection classifcation , if the model gives low precision then many comments will be said as aggressive, for high precision it will ignore the False positive values by learning with flase alarms. The precision can be calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Recall : Recall is opposite of Precision .When the model predicts. Precision used when the false Nagatives are high.In the aggressive detection classifcation problem , if the model gives low recall then many comments will be said as non aggressive, for high recall it will ignore the FalseNegative values by learning with flase alarms.The recall can be calculated as follows:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

F1 score: F1 score combines precision and recall and provides an overall measure of the models' accuracy. The value of the F1 score lies between 1 and 0. If the predicted value matches with the expected value, then the f1 score gives 1, and if none of the values matches with the expected value, it gives 0. The F1 score can be calculated as follows:

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{3}$$

4 Result And Discussion

Sub-Task A has been considered for classifying the aggression comments in social media. Since the dataset is imbalanced, several data preprocessing methods is adopted before classifying the aggression detection. Data augmentation with two types, machine translation, and noise addition, are used to resolve the imbalanced data issues. Finally, a fully machine-translated English data has been created to check the performance of existing deep learning models on machine-translated data that contains most noises. We have trained five deep learning models: LSTM, BiLSTM, LSTM-autoencoder, Word2vec, BERT transformer, and GPT-2 models using all datasets. The models are evaluated using evaluation metrics such as F1 score, accuracy, precision, and recall. We derived the metric evaluation result over the unseen test dataset. The performance of the models are shown in Table 2 and 3. The classification results for Autoencoder, LSTM, BiLSTM, word2vec and Bert models using raw data , raw data with augmented data and machine-translated English data are shown below:

TABLE-4 :Raw Data Classification results of different architectures on Subtask:A test data

Models	Set	Accuracy	precision (Weighted Average)	Recall (Weighted Average)	F1 Score (Weighted Average)
Autoencoder	English	0.78	0.68	0.78	0.70
	Bangla	0.55	0.47	0.55	0.41
	Hindi	0.58	0.52	0.58	0.47
LSTM	English	0.78	0.62	0.78	0.69
	Bangla	0.55	0.30	0.55	0.39
	Hindi	0.58	0.34	0.58	0.43
BiLSTM	English	0.68	0.63	0.68	0.65
	Bangla	0.57	0.45	0.57	0.50
	Hindi	0.56	0.45	0.56	0.50
Word2vec	English	0.78	0.68	0.78	0.72
	Bangla	0.61	0.59	0.61	0.55
	Hindi	0.64	0.60	0.64	0.60
BERT	English	0.79	0.79	0.79	0.79
BERT MultiLingual	Bangla	0.72	0.71	0.72	0.72
BERT Multilingual	Hindi	**0.69**	**0.69**	**0.69**	**0.69**
gpt2	English	**0.80**	**0.76**	**0.80**	**0.77**
gpt2	Bangla	**0.73**	**0.74**	**0.73**	**0.73**
gpt2	Hindi	0.63	0.62	0.63	0.62

TABLE-5:Raw data with Augmented Data Classification results of different architectures on Subtask:A test data

Models	Set	Accuracy	precision (Weighted Average)	Recall (Weighted Average)	F1 Score (Weighted Average)
Autoencoder	English	0.66	0.67	0.66	0.67
	Bangla	0.55	0.48	0.55	0.45
	Hindi	0.52	0.47	0.52	0.48
LSTM	English	0.60	0.67	0.60	0.63
	Bangla	0.54	0.47	0.54	0.49
	Hindi	0.49	0.47	0.49	0.47
BiLSTM	English	0.65	0.66	0.65	0.65
	Bangla	0.44	0.48	0.44	0.42
	Hindi	0.46	0.45	0.46	0.44
Word2vec	English	0.71	0.73	0.71	0.72
	Bangla	0.59	0.56	0.59	0.53
	Hindi	0.57	0.62	0.57	0.59
BERT	English	**0.75**	**0.78**	**0.75**	**0.77**
BERT MultiLingual	Bangla	**0.71**	**0.70**	**0.71**	**0.70**
BERT Multilingual	Hindi	**0.68**	**0.69**	**0.68**	**0.68**
gpt2	English	**0.75**	0.71	**0.75**	0.73
gpt2	Bangla	0.66	0.64	0.66	0.64
gpt2	Hindi	0.63	0.62	0.63	0.62

TABLE-6 :Machine translated English Data Classification results of different architectures on Subtask:A test data

Models	Accuracy	precision (Weighted Average)	Recall (Weighted Average)	F1 Score (Weighted Average)
Autoencoder	0.77	0.69	0.77	0.70
LSTM	0.74	0.64	0.74	0.69
BiLSTM	0.69	0.66	0.69	0.67
Word2vec	0.77	0.71	0.77	0.73
BERT	**0.78**	**0.78**	**0.78**	**0.78**
gpt2	0.76	0.76	0.76	0.76

We have observed that the gpt2 model performed best on English and Bangla raw data, whereas the BERT model performed best on Hindi raw data. Gpt2 gave the highest accuracy of 80 percent on English raw data. However, the Bert model performed best on the augmented and machine-translated datasets. It gave 0.78 percent accuracy on the machine-translated dataset. From the experiment, we have found that the BERT model performed best on noisy datasets, while the gpt2 model performed best on the raw dataset that does not contain any noise. It is aparent that, using machine translated data set is quite risky since the dataset contains noises and requires human intervention, which is costly and time-consuming. This work shows that the BERT model can work well on the noisy dataset. We evaluated the model with unseen raw data and got 78 percent accuracy, which is pretty good for industrial and future investigation purposes.

For the fully noisy dataset, the training-validation accuracy curve is repre- sented in Fig. 3. For all models, the training accuracy is higher than the validation accuracy, representing that the model has learned the dataset properly without being overfitted or under-fitted.

Figure 3 illustrates a high-level schematic representation of the model accuracy curve which is derived from BERT model using fully machine translated data .

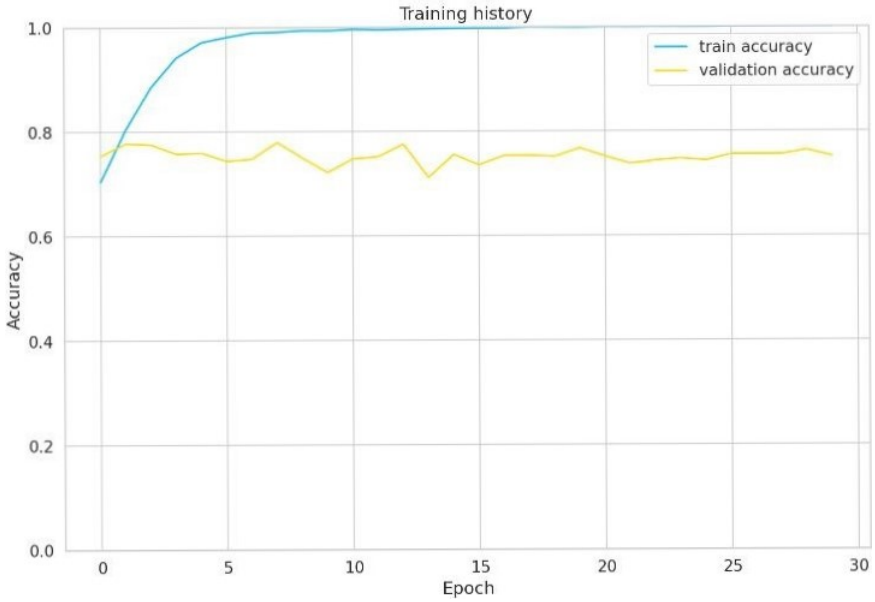

Figure 3: Curve of model accuracy derived from BERT model using fully machine translated dataset

The confusion matrix shows the number of True positive and False negative results has been predicted by each of the model. Figure 4 illustrates a high-level schematic representation of the Confusion matrix for Bert model using fully machine translated.

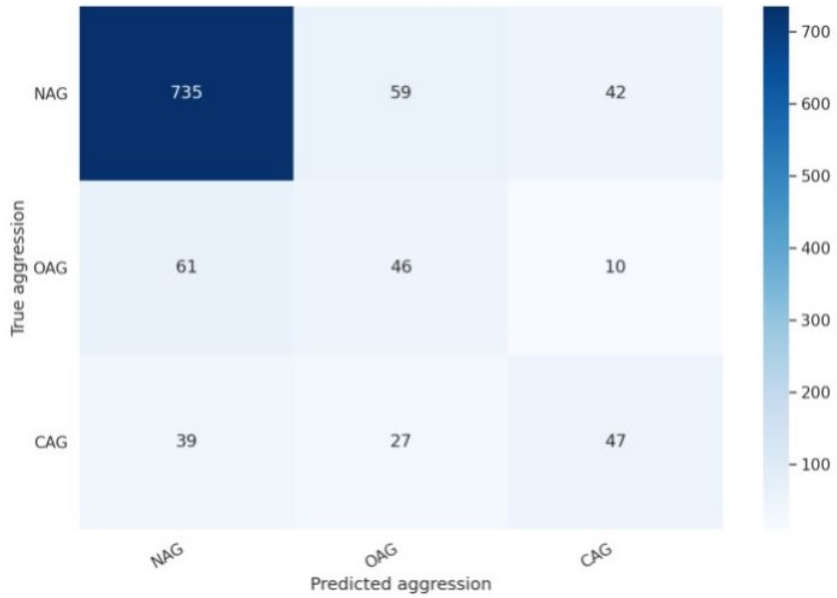

Figure 4: Confusion matrix for Bert model using fully machine translated dataset.

5 Conclusion

In this work, we have presented the process of generating machine-translated data and how the deep learning models perform on this dataset. We have made a comparative analysis with the performance of the models on raw data and semi-noisy datasets. The raw data denotes the dataset we have collected from the organization. The semi- noisy dataset denotes the augmented data we have added with raw data. We have used models such as LSTM, BiLSTM, LSTM-Autoencoder,word2vec, BERT, and GPT-2 and evaluated the performance of the models using performance metrics such as Accuracy, F1-score, Precision, and recall. The

Performance metric shows that the BERT model performed Best on the Machine translated and semi- noisy data, and the gpt2 model performed best on the raw dataset. The difference between the accuracy on raw, semi-noisy, and Machine translated is minimal. The highest accuracy for raw data is 80 percent, for semi-noisy data is 78 percent, and for machine-translated 78 percent. It is clear that training the BERT model using machine-translated data gives almost the same result as the raw dataset, which may be useful for the dataset that lacks a large dataset. Using our approach, future researchers will be able to analyze various problems associated with text datasets that were left behind due to the availability of dataset availability.

Acknowledgement

We acknowledge Mr. sabbir rahman for few important discussions.

Conflict of Interest

Authors of this article do not have any conflict of interest.

References

[1] R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri, "Benchmarking aggression identification in social media," in *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pp. 1–11, 2018.

[2] K. Kumari and J. P. Singh, "Ai_ml_nit_patna@ trac-2: Deep learning approach for multi-lingual aggression identification," in *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pp. 113–119, 2020.

[3] L. S. M. Altın, A. Bravo, and H. Saggion, "Lastus/taln at trac-2020 trolling, aggression and cyberbullying," in *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pp. 83–86, 2020.

[4] J. Lilleberg, Y. Zhu, and Y. Zhang, "Support vector machines and word2vec for text classification with semantic features," in *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pp. 136–140, IEEE, 2015.

[5] L. Wensen, C. Zewen, W. Jun, and W. Xiaoyi, "Short text classification based on wikipedia and word2vec," in *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, pp. 1195–1200, IEEE, 2016.

[6] E. M. Alshari, A. Azman, S. Doraisamy, N. Mustapha, and M. Alkeshr, "Improvement of sentiment analysis based on clustering of word2vec features," in *2017 28th international workshop on database and expert systems applications (DEXA)*, pp. 123–126, IEEE, 2017.

[7] T. Ranasinghe and M. Zampieri, "Multilingual offensive language identification with cross-lingual embeddings," *arXiv preprint arXiv:2010.05324*, 2020.

[8] F. Ramiandrisoa and J. Mothe, "Irit at trac 2020," in *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pp. 49–54, 2020.

[9] S. K. Tawalbeh, M. Hammad, and M. Al-Smadi, "Keis@ just at semeval-2020 task 12: Identifying multilingual offensive tweets using weighted ensemble and fine-tuned bert," *arXiv preprint arXiv:2005.07820*, 2020.

[10] H. Liu, P. Burnap, W. Alorainy, and M. Williams, "Scmhl5 at trac-2 shared task on aggression identification: Bert based ensemble learning approach," 2020.

[11] S. Tawalbeh, M. Hammad, and A.-S. Mohammad, "Saja at trac 2020 shared task: Transfer learning for aggressive identification with xgboost," in *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pp. 99–105, 2020.

[12] M.-A. Tanase, G.-E. Zaharia, D.-C. Cercel, and M. Dascalu, "Detecting aggressiveness in mexican spanish social media content by fine-tuning transformer-based models,"

[13] S. Bhattacharya, S. Singh, R. Kumar, A. Bansal, A. Bhagat, Y. Dawer, B. Lahiri, and A. K. Ojha, "Developing a multilingual annotated corpus of misogyny and aggression," *arXiv preprint arXiv:2003.07428*, 2020.

[14] D. Mandic and J. Chambers, *Recurrent neural networks for prediction: learning algorithms, architectures and stability*. Wiley, 2001.

[15] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based lstm for aspect-level sentiment classification," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 606–615, 2016.

[16] H. Nguyen, K. P. Tran, S. Thomassey, and M. Hamad, "Forecasting and anomaly detection approaches using lstm and lstm autoencoder techniques with the applications in supply chain management," *International Journal of Information Management*, vol. 57, p. 102282, 2021.

[17] K. W. Church, "Word2vec," *Natural Language Engineering*, vol. 23, no. 1, pp. 155–162, 2017.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[19] K. Lagler, M. Schindelegger, J. Böhm, H. Krásná, and T. Nilsson, "Gpt2: Empirical slant delay model for radio space geodetic techniques," *Geophysical research letters*, vol. 40, no. 6, pp. 1069–1073, 2013.

[20] M. C. Chen, R. L. Ball, L. Yang, N. Moradzadeh, B. E. Chapman, D. B. Larson, C. P. Langlotz, T. J. Amrhein, and M. P. Lungren, "Deep learning to classify radiology free-text reports," *Radiology*, vol. 286, no. 3, pp. 845–852, 2018.

[21] O.-M. Sulea, M. Zampieri, S. Malmasi, M. Vela, L. P. Dinu, and J. Van Genabith, "Exploring the use of text classification in the legal domain," *arXiv preprint arXiv:1710.09306*, 2017.

[22] Z. I. Attia, P. A. Noseworthy, F. Lopez-Jimenez, S. J. Asirvatham, A. J. Deshmukh, B. J. Gersh, R. E. Carter, X. Yao, A. A. Rabinstein, B. J. Erickson, *et al.*, "An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction," *The Lancet*, vol. 394, no. 10201, pp. 861–867, 2019.

[23] M. Heikal, M. Torki, and N. El-Makky, "Sentiment analysis of arabic tweets using deep learning," *Procedia Computer Science*, vol. 142, pp. 114–122, 2018.