International conference on Advanced Techniques in Communication Networking and Automation

RESEARCH ARTICLE OPEN ACCESS

A Deep Learning Based Integrated Approach to Detect Abusive Language in Social Media Text And Memes

Archana. B[1], Kesavarthini.C [3], Dr.K.Krishnakumari [3]

Students[1],[2]
Associate Professor [3]
A.V.C College of Engineering Mannampandal,
Mayiladuthurai.

ABSTRACT

In this paper an automated internet accessibility is system that can identify increasing facilitated and encouraged media,In the task, data freely expressed their opinions. However, sets are provided it also creates a space contaminators which is SemEval, OLIDwith of content to spread the attack uploaded to accomplish the tasks, We employed post like memes, comments, and tweets. Most of these attacks posts are three deep learning models which is RNN, LSTM, BERT techniques. Written between language and can easily evadeonline surveillance sytems.

Keywords:- LSTM, RNN, BERT, Deep Learning, Abusive Language Detection

I. INTRODUCTION

Social media is widely available and has grown to be a popular medium for users to communicate with one another. The accessibility of information in a multitude of formats is one of the factors contributing to these platforms' appeal. Social media has no content restrictions, so anyone can publish anything. Negative remarks based on things like ethnicity, gender, and religion are posted as a result. Thus, it is crucial to keep an eye out for and filter out any hostile or insulting language before it is published. Our project was focused on creating an automated system to recognize submitted comments and categorize them as hateful or not. Deep learning algorithms canbe used for the classification of comments.

II. RELATED WORK

The system was developed using machine learning, especially deep learning-based neural networks, which are learning algorithms based on artificial intelligence that can be created to recognize words, phrases, and sentences from all forms of communication, including websites, social media, and others. A large dataset of various words, phrases, and sentences that are classified as hateful, racist, bullying, discriminatory, etc. is used to train the system as the first step in creating an intelligent system. The following approach is suggested for this system: The dataset is initially split into two datasets (one containing the positive and negative remarks and the other containing the remaining classes). The datasets undergo preprocessing to clean the data and carry out additional processes, like labeling and removal of symbols. The information is then handed off to a subsequent stage, where text vectorization is carried out. The phrases are tokenized in this step. Following that, the data is split into training and validation groups and the text characteristics are extracted. The classifiers are given the training data, and their task is to place the data into one of the seven groups. The entire system was intended to be fully autonomous, have exceptional accuracy, and be implementable in real-world systems. In order to host large amounts of memory and provide access to computing capacity for real-time processing, the system can then be

connected to a cloud-based service. In Figure 1, the suggested method is displayed.

The dataset is split into two groups in the first step, one for the six categories of abusive language and the other for the positive and negative comments. A preprocessing stage is performed on the merged dataset, which involves data cleaning, symbol removal, labeling, etc. The second step, which involves text vectorization or word tokenization, receives the dataset after that. The extremely crucial stage of text feature extraction is then finished. With 80% training and 20% validation, the data are then divided into two groups for training and validation. Additionally divided into 80% validation and 20% proof are the 20% validation data. The classifiers are then trained to categorize text into one of the mentioned seven classes.

[1] The Sadiq et al. Several models were created and analyzed by Sadiq et al. in order to determine tweeting online abuse. The Multi-Layer Perceptron (MLP) with TF-IDF characteristics is one of the models.two deep neural networks, MLP, and word embedding CNN with LSTM and CNN with BiL-STM are the two networks. Results showed that using the TF-IDF and MLP With an accuracy of 0.92, the features-based model did better than other models.

2] Tamil. A multimodal framework (MemSem), suggested by Pranesh and Shekhar in 2020, consists of For image characteristics, use VGG19, and for text features, use BERT. MemeSem outperformed everyone with its outcome.multimodal baselines with 67.12% and unimodal baselines accuracy.

Furthermore, the advancements of language modeling and machine learning techniques show promising results when tackling the problem of offensive language identification. In the first edition of OffensEval [3], transfer learning methods such as BERT [4] proved to be among the most effective and accurate (as shown by the best system at OffensEval 2019 [6]) especially when dealing with limited labeled datasets. Still, some other models proved their predictive power as well, e.g., the C-BiGRU model which combines a Convolutional Neural Network (CNN) with a bidirectional Recurrent Neural Network (RNN) [7] that scored in the 9th position of OffensEval 2019.

In Reference[5] In order to detect offensive language, Mohaouchane et al. (2019)investigated the use of various Deep Learning architectures. On the dataset proposed in [12], several models, including CNN-LSTM, CNN-BiLSTM with attention, Bi-LSTM, and CNN model, were trained using AraVec embeddings of each comment.The CNN model was found to provide the best F1 score.

In Mubarak and Darwish [9] 36 million tweets were collected and used it to train a FastText deep learning model and SVM classifier on character n-gram features where it was found that the Arabic FastText DL model provided the

best results

For their proposed system, the reference of [10] gathered 17,567 Facebook posts that were annotated as having no hatred, little hate, or a lot of hate. They have a phrasethey like to use and a term they like. Support classifying linguistic texts.On the other hand, the longer sentences that were not observably accurate were broken down using the LSTM system, which was used to extend the ranges.vector machines (SVM) and long short-term memory (LSTM), a recurrent neural network, were the two techniques used in this research. (RNN).

In [16], the authors developed a dataset for Arabic speech collected from various sources, such as Facebook, Instagram, YouTube, and Twitter. They collected a total of 20,000 posts, tweets, and comments. They tested the dataset with 12 machine learning algorithms and two deep learning algorithms. They reported that the highest accuracy of 98.7% was achieved using the RNN.

III Proposed methodology

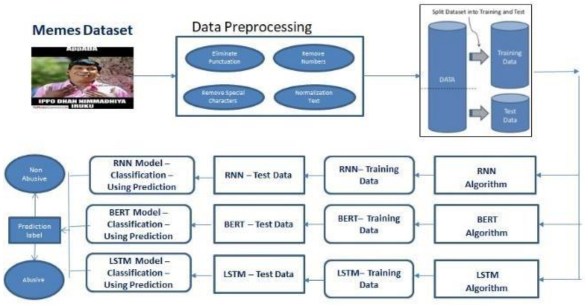

Figure 1 :Architecture of proposed model

Dataset:

The dataset will be used as training the dataset for the training of our model. The dataset was not balanced as it has fewer records of hate speech only as compared to overall comments. We use the SemEval,OLID dataset text abusive detection of our project. We use the own image datasets are used in our project.The comments in the dataset cannot be directly fed into the models constructed. So some pre-processing works were done.Pre processed comments were fed as input to the models after embedding. Embedding was done by using features such as Word2Vec, CBow, GloVe, Doc2Vec.We have used the following models for classification and results of these were compared.

RNN:

Recurrent Neural Network The RNN is an extension of a deep neural model that has ability to handle variable-length sequence input. Instead of learning features by the traditional feedforward structure, the RNN involves recurrent units which can use the information of the previous states. In addition, this architecture can effectively address the issue that the input of text content is in fixed size. Figure 5 illustrates the basic RNN framework and unfolded into a graph of the timesteps at time t. The input x_t is fed to the model at timestamp t, S_t is the hidden layer that captures input information x_t and previous state S_{t-1} at timestamp t−1, o_t illustrates the output of the model.

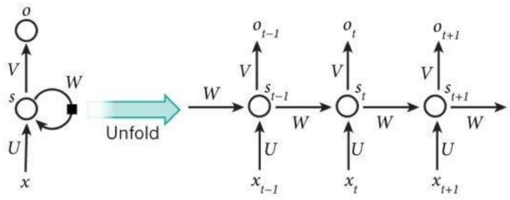

Figure 2 : RNN

BERT:

Google research team releases Bidirectional Encoder Representation from Transformer (BERT)and achieve state of the art results on many NLP tasks. BERT uses identical multi-head transformer structure that is introduced in (Vaswani et al., 2017). The model is pre-trained on huge corpus from different sources. Since the dataset size in this SemEval-2019 Task 6 is not that big, we pass the dataset into the pretrained BERT model, and report the loss and accuracy at each epoch. The observation from experiments shows that after 1st or 2nd epochs, the model converges fast and always get very lower loss on the validation set. In such case, in the sub-task B and sub-task C, we report the macroF1 score after the model trains after 1st, 2nd and 3 epochs.

We train the LSTM model using early stopping with a patience of five epochs over the validation loss. For Level A, we use an LSTM model with a hidden size of 128, a dropout rate of 0.3, a batch size of

256, and a learning rate of 0.0002. For Level B, the LSTM model has a hidden size of 50, a dropout rate of 0.1, batch size of 32, and a learning rate of 0.0001. Finally, the Level C LSTM model has hidden size of 50, a dropout rate of 0.1, batch size of 32, and a learning rate 0.0001. We use the Adam optimizer for training.

LSTM:

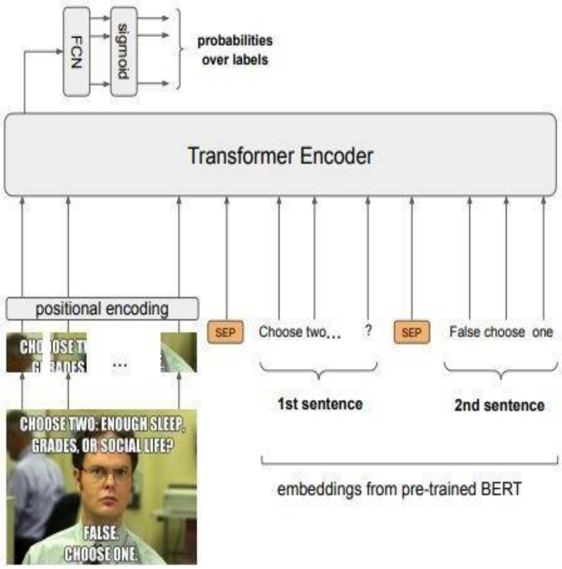

Figure 2 : BERT

To detect Abusive Language in multimodal memes. We first take advantage of the memes' visual components and create a number of RNN structures. The textual data is then taken into account, and deep

The LSTM model builds a continuous representation of the input tweet in a sequential manner, where at each step, it decides which information to update, reset, and output to the next step. The LSTM model can account for long-distance relations between words, but its loss of information along the steps can become severe. One partial solution of the information bottleneck problem, which we also use, is to use an attention mechanism, thus allowing for lookups of previous step outputs.

The first layer of the LSTM model is an embedding layer, which we initialize with a concatenation of the GloVe 300-dimensional and FastText's Common Crawl 300-dimensional embeddings. The embedding layer is then followed by a dropout and a bi-directional LSTM layer with an attention mechanism on top of it. We concatenate the results of the attention mechanism with both averaged and maximum global poolings over the outputs of the LSTM model. The final prediction is produced by a sigmoid layer for Levels A and B, where we have binary classification, and softmax for Level C, where we have three classes. learning-based approaches (RNN, LSTM,BERT) are used for categorization. Both textual and image input are taken into the classifier using Deep Learning approaches such as BERT+RESNET.

BERT+RESNET:

We designed a multi-task learning model that can automatically generate sub-labels.A supervised approach to detect multimodal hateful memes, the primary multimodal task consists of using BERT and RESNET to extract features Two identical subtasks share the underlying feature learning network in the hard state Parameter distribution method. The primary task part is a multimodal classification net, which consists of three steps,the extraction of features, the fusion offeatures, and the output of classification. Pre-trained models have performed very well in recent years, so we used two pre-trained models as the backbone for two unimodal tasks in the hateful memes detection task.

1 1

twelve-layers BERT to extract text feature Ft.

$$F_t = \text{BERT}\ (I_t; \theta_t^{\text{bert}}),$$

where It is the text input, θ_t^{bert} is all parameters of the BERT we used.

RESNET101 [40] to extract image feature Fv

$$F_v = \text{RESNET}\ (I_v; \theta_v^{\text{resnet}}),$$

where Iv is the image input, θ_v^{resnet} is all parameters of the RESNET we used.

Then, the text and image representations are concatenated as Fm = [Ft; Fv] and projected onto a low-dimensional space.

$$F_m^* = \sigma(W_1^m F_m + b_1^m),$$

where W_1^m and b_1^m are the parameters of the first linear layer in the primary multimodal task, σ is the activation function. After that, we use the representation of fusion obtained from the linear layer and activation function to detect whether the meme is hateful.

$$\hat{y}_m = W_2^m F_m^* + b_2^m,$$

Where,$W_2 \in R^{d_m \times 1}$, and W_2^m and $b_2^m$are

For text processing, we use the pre-trained

Where, $u \in \{t, v\}, W^u$ and b^u are parameters of the first linear layer in the unimodal auxiliary task.

Then, the results of unimodal auxiliary tasksare obtained by

$$z_u = W_2^u F_u^* + b_2^u,$$

Where, $u \in$ $\{t, v$

For image processing, we use the pre-trained

of the second linear layer in the unimodal auxiliary task.

IV. RESULT AND EVALUATION TECHNIQUE:

For English :

Hence the dataset is not balanced we use the performance metrics Precision, Recall, F1

1 1 score.

Table 1.1 Result of abusive Language detection across 3 datasets. The Result are shown as precision, recall, f1

MODEL	PRECISION	RECALL	F1
RNN	0.6447	0.6083	0.8000
LSTM	0.9160	0.90	0.9160
BERT	0.90	0.88	0.890

the parameters of the second linear layer in the multimodal primary task.

The auxiliary tasks are two unimodal classification tasks that detect the presence of hateful sentiment in text and images, respectively. We project the unimodal features into a new feature space, which reduces the impact of the dimensional difference between different modalities. Moreover, the text and image auxiliary classification tasks share modal features with the primary multimodal classification task.

$$F_u^* = \sigma(W_1^u F_u + b_1^u),$$

CONCLUSION

In this paper, We present social media is an important way of communicating on social media platforms like Twitter, facebook,youtube, etc,.peoples are posting their opinions that have can impact on a lot of users. The comments and memes that contain positive, negative and mixed feeling of words as well as memes and the comments contain abusive and not abusive words are classified as abusive language identification. For identifying on social media text and memes in abusive language in different pretrained models like, RNN, BERT, LSTM are text feature extraction, RESNET+BERT are visual feature extraction. Among the obtained result for identiyfing abusive text and memes for BERT model accuracy is 0.74 and LSTM model accuracy is 0.73.

REFERENCES

[1].Saima Sadiq, Arif Mehmood, Saleem Ullah, Maqsood Ahmad, Gyu Sang Choi, and Byung-Won On. 2021. Aggression detection through deep neural models on twitter. Future Generation Computer Systems,114:120–129.

[2] Raj Ratn Pranesh and Ambesh Shekhar. 2020. Meme-sem:a multimodal framework for sentimental analysis of meme via transfer learning.

[3] Marcos Zampieri et al. "SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)". In: Proceedings of the 13th InternType equation here.ational Workshop on Semantic Evaluation. 2019, pp. 75–86 (cit. on pp. 7, 18).

[4] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: NAACL-HLT 2019: Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2019, pp. 4171–4186 (cit. on p. 7).

[5] Mohaouchane, H., Mourhir, A., and Nikolov, N. S. (2019).Detecting offensive language on arabic social media using deep learning. In 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), pages 466–471. IEEE.

[6] Ping Liu, Wen Li, and Liang Zou. "NULI at SemEval-2019 Task 6: Transfer Learning for Offensive Language Detection using Bidirectional Transformers". In: Proceedings of the 13th International Workshop on Semantic Evaluation. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019,pp. 87–91. doi: 10 . 18653 / v1 / S19 - 2011. url:https://www.aclweb.org/anthology/S19-2011 (cit. on p. 7).

[7] Jelena Mitrovi´c, Bastian Birkeneder, and Michael Granitzer. "nlpUP at SemEval-2019 Task 6: a deep neural language model for offensive language detection". In: Proceedings of the 13th International

Workshop on Semantic Evaluation. 2019, pp. 722–726 (cit. on p. 7).

[8] Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., and Patel,
A. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In Proceedings of the 11th Forum for Information Retrieval Evaluation, pages 14–17.

[9] Mubarak, H. and Darwish, K. (2019). Arabic offensive language classification on twitter. In the International Conference on Social Informatics, pages 269–276. Springer.

10.Del Vigna, F.; Cimino, A.; Dell'Orletta,F.; Petrocchi, M.; Tesconi, M. Hate me, hateme not: Hate speech detection on Facebook.In Proceedings of the First Italian Conference on Cybersecurity, Venice, Italy,17–20 January 2017; pp. 86–95 [11]Mubarak, H., Darwish, K., Magdy, W.,Elsayed, T., and AlKhalifa, H. (2020).Overview of osact4 arabic offensivelanguage detection shared task. 4.

[12] lakrot, A., Murray, L., and Nikolov, N. S. (2018). Towards accurate detection of offensive language in online communication in arabic. Procedia computer science, 142:315–320.

[13] Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. Joint modeling of emotion and abusive language detection. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4270–4279, Online, July 2020. Association for Computational Linguistics.

[14] Xianbing Zhou, Yang Yong, Xiaochao Fan, Ge Ren, Yunfeng Song, Yufeng Diao, Liang Yang, and Hongfei Lin. Hate speech detection based on sentiment knowledge sharing. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 7158–7166, 2021.

[15] Ziqi Zhang, David Robinson, and Jonathan Tepper. Detecting hate speech on twitter using a convolution-gru based deep neural network. In European semantic web conference, pages 745–760. Springer, 2018.sss

16. Omar, A.; Mahmoud, T.M.; Abd-El-Hafeez, T. Comparative performance of machine learning and deep learning algorithms for Arabic hate speech detection in osns. In Proceedings of the International Conference on Artificial Intelligence and Computer Vision, Cairo, Egypt, 8–10 April 2020; Springer: Cham, Switzerland, 2020; pp. 247–257.