ABUSIVE COMMENT CLASSIFICATION IN DIGITAL PLATFORMS

GOWTHAMI NAGAPPAN – B1326237@live.tees.ac.uk – Teesside University – UK
CIS4050-N-FJ1-2021 Deep Learning - School of Computing, Engineering & Digital Technologies

INTRODUCTION

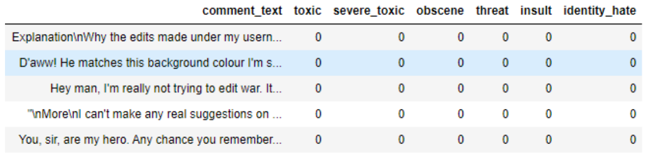
- In the current century, Digital communication platforms like Wikipedia, twitter, Glassdoor has created to learn and share many updated information as well as it has become a unique place for people to freely express their opinions. Meanwhile, there are some groups that are taking advantage of this framework and misuse this freedom to implement their toxic mindset (i.e., insulting, verbal sexual harassment, threads, Obscene, etc.).
- 73% of the adult internet users have seen someone be harassed in some way online, 40% of the internet users have personally experienced online harassment, and 45% of those have experienced severe harassment. Sometimes extreme cases of cyber-bullying even lead the victims to commit suicide.
- This Challenges are solved by LSTM Model by classifying comments with high accuracy by pertaining models with larger length of text.

DATASET

- The dataset is from Kaggle's online toxic comment dataset (jigsaw toxic comment classification challenge).
- The dataset spitted into 159570 Train data and 153163 Testing data.

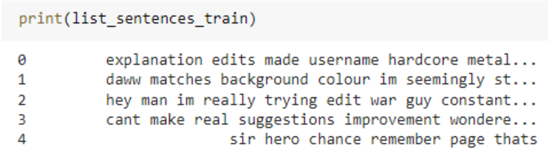
METHODOLOGY

DATASET



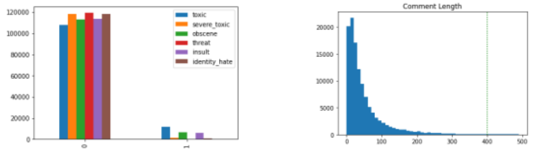
comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
Explanation\nWhy the edits made under my usern...	0	0	0	0	0	0
D'aww! He matches this background colour I'm s...	0	0	0	0	0	0
Hey man, I'm really not trying to edit war. It...	0	0	0	0	0	0
"\nMore\nI can't make any real suggestions on ...	0	0	0	0	0	0
You, sir, are my hero. Any chance you remember...	0	0	0	0	0	0

DATA PRE-PROCESSING



```
print(list_sentences_train)

0        explanation edits made username hardcore metal...
1        daww matches background colour im seemingly st...
2        hey man im really trying edit war guy constant...
3        cant make real suggestions improvement wondere...
4                      sir hero chance remember page thats
```

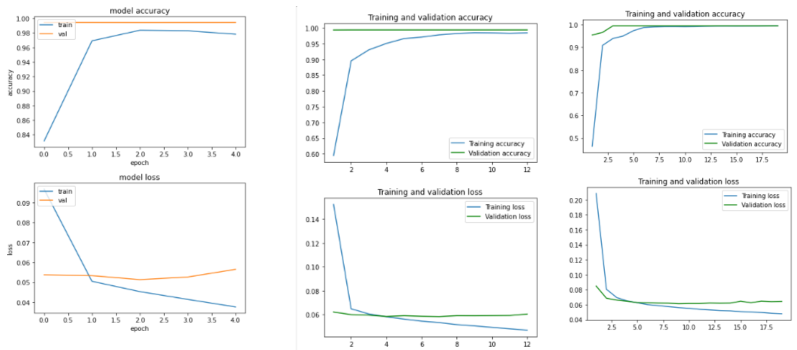
DATASET VISUALIZATION



EXPERIMENT

PARAMETER TURNING METHOD

1.) LSTM 2.) LSTM_Small_25_20_6 3.) LSTM_Tiny_20_10_6

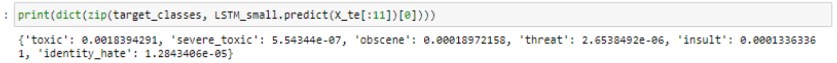


No improvement By
Reducing Learning Rate.
Dropout , Early Stopping Methods

Still Overfitting: Training Accuracy – 0.9894
Validation Accuracy - 0.9942
T.loss – 0.0496 / V.loss – 0.0625

Final Accuracy : 0. 9840
0.9942
0.0479 / 0.0644

RESULT :



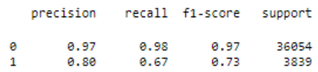
```
: print(dict(zip(target_classes, LSTM_small.predict(X_te[:11])[0])))

  {'toxic': 0.0018394291, 'severe_toxic': 5.54344e-07, 'obscene': 0.00018972158, 'threat': 2.6538492e-06, 'insult': 0.0001336336
  1, 'identity_hate': 1.2843406e-05}
```

CHALLENGES

Sequence Learning Problem
Value Memorization
Echo Random Integer
Echo Random Subsequence
Sequence Classification

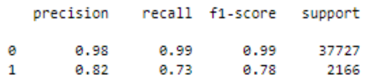
EVALUATION METRICS

Toxic :



```
   precision    recall  f1-score   support

0       0.97      0.98      0.97     36054
1       0.80      0.67      0.73      3839
```

Obscene :



```
   precision    recall  f1-score   support

0       0.98      0.99      0.99     37727
1       0.82      0.73      0.78      2166
```

CONCLUSION

As result from above 3 model, the considerable accuracy was achieved from the 3rd LSTM Model of parameter tuning with the weight features of 20_10_6 compare rest of 2 models. also, the performance of this model has good accuracy in evaluation metrics. So, model with less learning capacity gives a good result in in large dataset.

ETHICAL CONCERNS

Data Bias
Data Interpretation
Language Barrier

FUTURE WORK

- I suggest to develop a LSTM Model in future to detect and classify the toxic words in Images and videos in all digital platform.
- The user should get warning like “your comments like abusive / are u sure to post ”.Classification in quick span of time in texts, images and videos before posting into any platform.

REFERENCE

- Toxic Comment Classification Challenge: Identify and classify toxic online comments(2018).Accessedfrom https://www.kaggle.com/datamunge/sign-language-mnist
- Georgakopoulos, S.V., Tasoulis, S.K., Vrahatis, A.G., Plagianakos, V.P.: Convolutional neural networks for toxic comment classification. In: Proceedings of the 10th Hellenic Conference on Artificial Intelligence, pp. 1–6 (2018)

CIS4050-N-FJ1-2021 - DEEP LEARNING

CLASSIFICATION OF ABUSIVE COMMENTS IN SOCIAL MEDIA USING DEEP LEARNING

SCHOOL OF COMPUTING, ENGINEERING AND DIGITAL TECHNOLOGIES

MAY 6, 2022

Prepared by

B1326237 - GOWTHAMI NAGAPPAN

ABSTRACT

Social media platforms are used by billions of users worldwide on a daily basis. Safer usage of the platforms is very important and the companies should invest enough resources to achieve the objectives. The research paper [1] discusses about the various deep learning algorithms which are used to train the models to classify the comments. The main findings of the research paper is to determine whether the comments are abusive or not using different algorithms. Automatic detection of these comments are segregated into various categories which are then filtered. The algorithms are also compared with one another to find their effectiveness and the accuracy levels are also monitored. The paper concludes saying that CNN with Glove algorithm is very effective in comparison with their peers.

INTRODUCTION

Social media is used by all sort of people to interact with each other and post their views or opinions where the people can view and make comments on it. These posts might have an impact on certain section of people or community based on the vulgarity.

Safer usage of the platforms is very important and the companies should invest enough resources to achieve the objectives. A report which was taken in Jan 22 states that there are 4.62 billion users around the world who are using social media platforms. On an average 15% of the people using the platform are facing some type of online Harassment which are physical threats, stalking, Identity hate and sexual harassment. Avoiding abusive comments on the platform is a huge task and several measures which are taken until now has many pitfalls due to the timeline it is taking to intervene and detect them.

In this paper, the research has been carried to mitigate these challenges and how deep learning can be used to facilitate the process of detecting the abusive comments within a short period of time. On various platforms, they are still monitored manually and it is taking huge resources for the social media companies. Instead of using the older process, the latest technologies can be implemented which can help the firm to reduce the dependence on manpower and utilise the resources effectively. A study by ‘pew’ suggests that the younger generation getting exposed to the toxic posts and vulnerabilities is very high which can lead to severe problems.

METHODOLOGY

The first methodology, Convolutional Neural Network - CNN is used in the paper to categorize the text. In Natural language processing, each row of text is vectorized into fixed dimension. N-gram is generated over the words and to make the CNN model very effective, Glove embedding has been used and the comments were padded to have a similar size of input and output. Later with a max-pooling and maximum activation value, the result is passed through the whole text. Finally, the simple CNN model and Glove with CNN model was implemented to get the desired results.

The second methodology, LSTM (Long Short Term Memory) is better than RNN in terms of memory. It enables to classify, process and make predictions based on time series data. Pretrained Glove model are used with LSTM to classify the comments in a selected dataset.

Final deep learning method designed on CNN-LSTM with Glove is used to classify the Text into different categories.

EXPERIMENTS

Dataset

The dataset was taken from ‘Wikipedia talk page edits’ where 160,000 comments was sourced, and they are named under various categories. Some of them are part of more than one category.

Data Visualization

These multi class comments are visualized by bar chart and correlation matrix to understand the count of different comments according to different categories and correlation between each category.

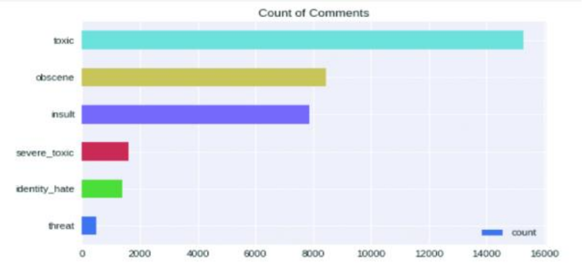


Fig: Bar Chart

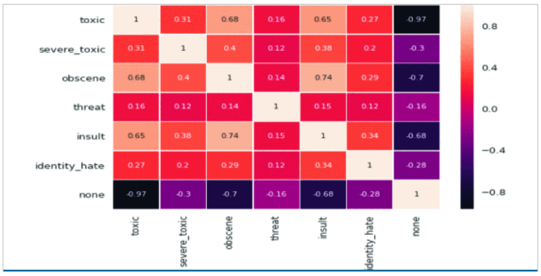


Fig: Corelation Matrix

RESULT AND DISCUSSION

The results on comparison of different algorithms are observed in the paper and the finding are mentioned below. Simple convolutional neural network training accuracy is 97.85, Validation accuracy is 98.01 and Testing accuracy is 97.06. When comparing with Glove & CNN, the simple CNN has achieved higher training accuracy, but the validation and Testing accuracy are a bit low. In the same way, when we compare simple LSTM against Glove & LSTM, the simple LSTM model accuracy looks better.

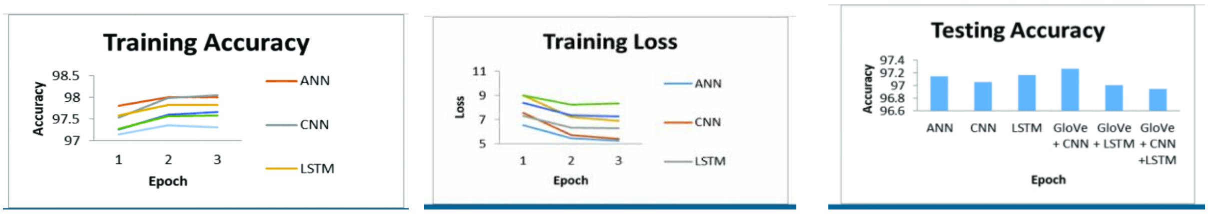


Fig: Training and Testing Result Graphs

Finally, when all the different models are compared with each other, CNN with Glove model performance was good. The results support the conclusions drawn in the paper. As per my understanding and dwelling deep into the research paper, there are no limitations which I have observed. But the only constraint which I can see is, the paper was not able to provide the detailed information on how the text is processed and the output is received.

RELATED WORK

The Previous related work on abusive comment classification done by Dawie Yin and his colleagues tried a context-based approach in 2009 [2] and then few machine learning algorithm was used to analyse the sentiment from the text. Later on, these methods were improved which was described on the paper, "Using crowdsourcing to improve profanity detection" to filter spam and blacklist them. The research document, "Abusive Language Detection in Online User Content" [3] was published by Yahoo which was taken as a reference. All the related research papers have concluded the accuracy of comment classification. These approaches are taken as reference to implement these techniques using deep learning techniques.

CONCLUSION

In this research paper, a model is trained with different deep learning algorithms to classify the multi-labelled comments in Digital platforms with various categories like toxic, obscene, threat, insult, severe toxic and identity hate. This paper concludes stating that Glove & CNN performs the best when compared with other algorithmic techniques. In my part 1 poster work, I have tried with simple LSTM and Glove with LSTM to classify the Multi class comments.

According to part 1 practice, the CNN Convolutions and pooling technique lost information of order of the words, so that Sequence extraction was tough to fit into the CNN architecture. From my practice, I would like to Conclude ‘LSTM with Glove’ Deep learning model can be trained with a large number of datasets with long length sequence which will work better when compared to CNN and other models like RNN. Adjusting Few parameters and early stopping help to achieve a good accuracy with desired dataset. The text was classified into several abusive categories which can be identified as a good result.

REFERENCE

1. Classification of Abusive Comments in Social Media using Deep Learning
   Authors : Mukul Anand; R. Eswari
   https://ieeexplore.ieee.org/abstract/document/8819734/references#references

2. Z. Xue Yin, L. Hong, B. D. Davison, A. Kontostathis and L. Edwards, "Detection of Harassment on Web 2.0. in CAW 2.0 '09", Proceedings of the 1st Content Analysis in Web 2.0 Workshop, 2009.

3. Sara Sood, Judd Antin and Elizabeth Churchill, "Using crowdsourcing to improve profanity detection", AAAI Spring Symposium: Wisdom of the Crowd, 2012.

4. Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad and Yi Chang, "Abusive Language Detection in Online User Content", Proceedings of the 25th International Conference on World Wide Web (WWW '16).

5. Abusive Comments Classification in Social Media Using Neural Networks
   D. R. Janardhana, Asha B. Shetty, Madhura N. Hegde, Jayapadmini Kanchan, and Anjana Hegde
   https://www.researchgate.net/publication/343392211_Abusive_Comments_Classification_in_Social_Media_using_Neural_Networks

CIS4050-N-FJ1-2021 DEEP LEARNING

SELF REFLECTIVE REPORT

Patience is a key aspect while learning neural network algorithms and computational models which is used to tackle deep learning problems. The inputs gained from the class and lab sessions gave me the ability to derive solution for the projects.

As a group, we did not have much idea about the project work at the initial stage. Starting from the selection of the project, the group mates were very rude with me and I was not comfortable from the begining. As I informed the same with you earlier, you advised me to do the project on my own. So, I started working on the project and I informed the group mates that I'm planning to work alone and do the presentation individually. Then they asked me about the project work and ideas and were interested to work with me as a group. So, we decided to do text classification in 'NLP' which was very harder and also a broader area to explore. After studying few research papers about 'LSTM' model in Text classification, I finalised the dataset which is related to a real-world example. As a group, we divided the work and a fixed time was allocated to propose a solution for the project.

Having a basic knowledge on the domain, it was hard to study and understand each concept of 'LSTM' multi-classification technique. Learning about several models and practicing each one of them helped me to improve the accuracy with a basic simple model. After successfully implementing the basic model, I started exploring few more concepts.

Working with this group was very tough and I requested them for online meeting but they never agreed and asked for a physical meeting. I went for in-person meeting for a couple of times and was able to observe the lack of knowledge with them. So, I shared few links and ideas so they can get a basic understanding on the concepts. But again, there was a severe lack of knowledge and they were just calling for group meetings. Also, I used to share the updates about the code and the poster in the WhatsApp group, but I don't get a response. Then I shared the poster to you and based on the recommendations I did the changes on both the coding and poster.

During the final week (11th) lab session, each group got 2 hours of time to concentrate on part-1 project work. During this time, my groupmates literally fighted with me for no reason and they said that they did some work. So, I asked them to share the code and screenshot of the result to add it in the poster and asked them to give feedback of the poster which I already created. They said to change the structure and colour. I changed accordingly and waited for the screenshot of the results so I can add it to the poster. I waited until 27th April but there was no update from them and finally, I asked them to add their work on the poster by themselves. On 2nd May they did few changes in the introduction part on the poster which I had created. 3rd May they shared a completely different poster with the wrong results and coding.

I advised them on the graph which was completely incorrect and they said that they were busy and can't do anymore changes. I was also not feeling well and it was difficult for me to change at the last minute. They called me on 4th may and said their poster is a final one and we are going to present it. They were not listening to my words and I had no option except to proceed with the same poster but on the other side I worked on it to make sure that I implement the right alogrithms.

They said that I am speaking on the methodology part and did not listen to me. Since it’s a group work, I was in the position to accept it.

The problem with the group was they wanted to dominate me and don’t want to use my inputs. I am so adjustable, but they tried to rule and scold me for no reason. In last, they confidently presented the work which was completely wrong.

If I worked with a right group, I would have given excellent inputs for the project where I would have gained much more knowledge on various topics. I sincerely apologize for the problem and I am attaching both the coding and poster work I did.

Working as part of a group, I learnt that I should have been in-flexible and very rigid on the subject matter so they would have adhered to the project procedure. This gave me a better learning experience on how to deal while working on the group.