CONTENT MODERATION USING DEEP LEARNING

PROBLEM STATEMENT

- There has been an enormous increase in objectionable content on different social media platforms. This daily onslaught of disturbing posts can lead to conflicts on the web. Such content needs to be pulled down from websites. This is where content moderation comes into the picture.
- The scale at which these platforms operate makes manual content moderation nearly impossible, leading to the need for automated or semi-automated content moderation systems.
- Instead of using different models for different platforms a cross-platform analysis would be an effective solution.

STATE OF THE ART

Research Papers :

- Deep Learning for Hate Speech Detection in Tweets by Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma.
- Hateful symbols or Hateful People? Predictive Features for hate speech detection on Twitter **by** Zeerak Waseem and Dirk Hovy .
- Convolutional Neural Networks for Toxic Comment Classification by Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis, and Vassilis P. Plagianakos .
- Aggression Identification using Deep Learning and Data Augmentation by Julian Risch and Ralf Krestel .

OBJECTIVE

- We need to create a model which works best on the platform for making automated content moderation systems.
- A cross-platform analysis using a co-trained model would be a cost-effective solution.

DATASETS

- **Twitter 1**
- **Twitter 2**
- **Quora**
- **Reddit**

METHODOLOGY

- **Data Collection:**
  - Twitter : collection of data is done through Twitter streaming API.
  - Reddit : Collection of data is done using pushShift .
  - Quora : Data is taken from Kaggle.
- **Data Preprocessing:**
  - Removal of Punctuation marks , special characters, urls etc.
  - Normalising the text.
  - Tokenization of sentences.
  - Removal of stopwords.
  - Lemmatization of words

- **Training and Testing**
  - **Train_test_splits:** scikit-learn provides a helpful function for partitioning data, train_test_split , which splits out your data into a training data of about 50% and the remaining data for validation and testing each equal in size.
- **Evaluation**
  - Following deep learning models have been used for evaluation:
    - LSTM
    - GRU
    - MLP
    - CNN
    - RNN

RESULTS AND ILLUSTRATIONS

Evaluation Metrics

- Accuracy
- Precision
- Recall
- F1-score

MODEL	FEATURE INPUT	ACCURACY	PRECISION	RECALL	F1 SCORE
CNN	Word2Vec	79.2%	85.2%	29.2%	43.5%
	Glove	83.3%	82.1%	50.2%	62.3%
	FastText	80.4%	85.3%	34.6%	49.2%
MLP	Word2Vec	77.8%	85.1%	23.1%	36.3%
	Glove	80.9%	73.3%	47.8%	57.9%
	FastText	77.9%	72.2%	31.5%	43.9%
LSTM	Word2Vec	79.6%	91.0%	28.4%	43.3%
	Glove	84.0%	77.7%	58.4%	66.7%
	FastText	80.5%	76.2%	41.9%	54.0%
GRU	Word2Vec	80.3%	84.1%	34.7%	49.1%
	Glove	84.6%	83.5%	54.6%	66.0%
	FastText	81.1%	90.8%	34.6%	50.1%
RNN	Word2Vec	77.1%	89.0%	16.0%	35.8%
	Glove	81.5%	77.7%	45.8%	57.6%
	FastText	79.2%	82.3%	29.5%	48.0%

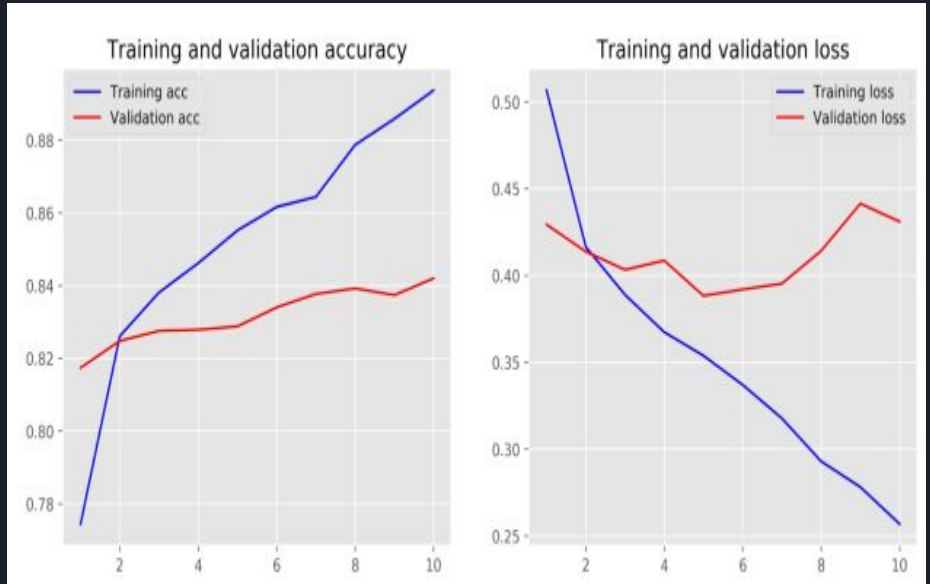

Fig 1.2 Accuracy and Loss graphs for training and validation (Model - LSTM, Feature Input-Glove)

CONCLUSION

In this project ,we have evaluated a number of deep learning approaches and identified those most suitable for detecting sensitive content. We have shown it is possible to analyse a large body of social media data using deep learning in a reliable and replicable way by employing a methodology to collect, train and classify data.We also compared the accuracy of different deep learning models and found out that **LSTM performs best using GloVe word embeddings** on our dataset.

PROPOSED WORK PLAN FOR FUTURE

We would further apply all the deep learning models on the rest of the platforms and would perform a multi-platform comparison and analyze which model best suits the platform. Also, we would perform a cross-platform content-moderation using a co-trained model enriched with a small amount of labeled data from a different platform, to moderate content on that platform. A multimodal approach is used for image and text datasets. Finally, we would deploy a portal as a proof-of-concept for different methods.

THANK YOU