Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019)
IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4

Classification of Abusive Comments in Social Media using Deep Learning

Mukul Anand
Department of Computer Application
National Institute of Technology
Tiruchirappalli, Tamil Nadu, India

Dr.R.Eswari
Department of Computer Application
National Institute of Technology
Tiruchirappalli, Tamil Nadu, India

Abstract— Social media has provided everyone to express views and communicate to masses, but it also becomes a place for hateful behavior, abusive language, cyber-bullying and personal attacks. However, determining comment or a post is abusive or not is still difficult and time consuming, most of the social media platforms still searching for more efficient ways for efficient moderate solution. Automating this will help in identifying abusive comments, and save the websites and increase user safety and improve discussions online. In this paper, Kaggle's toxic comment dataset is used to train deep learning model and classifying the comments in following categories: toxic, severe toxic, obscene, threat, insult, and identity hate. The dataset is trained with various deep learning techniques and analyze which deep learning model is better in the comment classification. The deep learning techniques such as long short term memory cell (LSTM) with and without word GloVe embeddings, a Convolution neural network (CNN) with or without GloVe are used, and GloVe pretrained model is used for classification

Keywords— CNN, LSTM, GloVe, RNN, BOW, TF-IDF, Embeddings.

I. INTRODUCTION

Today social media has become the place for any discussion to take place because of its reach and accesses ability. Social media has given power to every individual to express themselves in front of others, but this platform is becoming the platform for attacks on people based on characteristics like race, ethnicity, gender and sexual orientation, or threat of violence towards others. According to PEW Research Institute survey in 2014[1], 73% of the adult internet users have seen someone be harassed in someway 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 leads the victims to commit suicide. The social media platforms are keen determine the online abuse by the users report the abusive comments, but industries are also looking a way to automate this. Since the researchers in the field of natural language processing (NLP) and machine learning traditionally haven't applied their research for commenting spaces and abuse.

In this paper, deep learning is applied to determine, whether the comment is abusive or not and taking it to further classifying it to different category like toxic, severe_toxic, obscene, threat, insult, and identity hate. In this paper, we use simple neural network, convolution neural networks (CNN) and long short term memory (LSTM) with or without GloVe pretrained models. We use feature embeddings and LSTM to determine the most salient features of abusive comments. Lastly, we fine-tune the models before testing them on our datasets, and then compare these neural networks. The dataset which we use is comments from Wikipedia's talk page edits. The dataset initially have three categories with some improvements by The Conversation AI team, a research initiative founded by Jigsaw and Google. For training more than 160K comments and for testing the model 153K comments. We tested our models and used accuracy metrics, to know how well different model is working.

II. RELATED WORK

Since the severity of abusive comments in social networks are known and not much work has been done to prevent users from online social media abuse. But, there is an urgent need for a better system for detecting and barring these contents online.The earlier efforts on abuse classification goes back to 2009, where Dawie Yin and his colleagues explored a context based approach[2]. They have used content features, sentiment features and context features of a comment. They used a supervised machine learning approach, Support Vector Machine(SVM) with n-grams proves to be better than previous method TF-IDF. In 2012 S.O.Sood[3], came up with another approach which is inspired from commercial rule-based spam filtering using blacklists. They used blacklists along with an edit distance metric and showed that their approach is better for online profanity detection compared to which existed earlier.

Yahoo seems to automatically moderate online abuses, which they documented in a recent publication[4]. Yahoo used a combination of parser, lexical and syntactic features to train a

classifier using supervised machine learning algorithms. Google Jigsaw recently published a paper[5] that used data from Wikipedia detox project[6]. The paper discusses the effectiveness of logistic regression and multi-layer perceptron for abusive language classification. The paper also compared with human baseline. Method discussed in this paper used in google perspective API- an API that takes in a piece of text and returns that the text is abuse or not. All the published works mentioned above seems to focus more on the data through progressive feature selection. And there is not enough effort have been put to explore deep learning techniques to detect the abusive comments online. Deep learning methods often require little or sometimes no feature engineering. However there has been significant work has been done that employs deep learning for text but for completely different problems.

III. PROPOSED APPROACH

Abusive comments classification in social media networks

Past few years social media platforms have done in detecting abusive content to make it more secure place than ever, but still these platforms need to people to report for the content is abusive or not. And there is need to classify these comments further into different categories, for a better understanding of people behaviour online. Amir H. Razavi[9] and his colleagues classified comments into categories like racism, homophobia, extremism, etc. In this paper, we are classifying the social media comments into toxic, severe toxic, obscene, insult, identity hate, threat. We hope that these categories will help in classifying the comments in better way and stop the people who are using social media for wrong reason.

Classification using deep learning

Convolution Neural Network : Convolution Neural Networks(CNN), is mostly used in computer vision, but recently CNNs have provided ground breaking results in the field of Natural Language Processing. For each word there is a row of vectors of fixed dimension. By using convolution, n-gram can be generated just like a sliding window of different size passed all over the words.

Using property of GloVe that similar words have similar cosine distances and cosine distances are similar to dot products and the dot product is actually a convolution. From the pair of the word embeddings, convolutional filters will learn meaningful features. For text, 1-D convolution because we slide the window in only one direction. Padding is required, so that the size of the input and the output is the same. Apply max-pooling, taking the maximum activation value, output from the convolution passing through the whole text. Next apply more dense layers and multi-layered perceptron on the top of these features and train it for classification task.

Long Short Term Memory : Humans don't start to think from scratch every time they encounter something new, they understand these things based on past knowledge. And this where the traditional neural networks lacked. Recurrent neural network are networks with loops in them, which allows the information to persist. But, RNN has problem of Vanishing gradient and that can be solved by LSTM. LSTMs are special kind of RNN, which is also capable of long term dependencies. These also have a chain like structure instead of a single network layer. The core idea behind LSTMs:

The key of LSTM is the cell state, which run down the entire chain, with some interactions. It's ability to add or remove info to cell state is regulated by structures like gates.

1. forget gate: takes the input from previously hidden layer and output 0 or 1. 0 means forget and 1 means remember.
   $Ft = \sigma(Wf . [ht-1, xt]+ bf)$ (1)
2. Input Gate: decides what new information to update in cell state.
   It has two parts:
   - A sigmoid function which decides values to be updated.
     $It = \sigma (wi . [ht-1, xt] + bi)$ (2)
   - tanh function which creates a vector of new candidate values.
     $Ct = \tanh (Wc . [ht-1, xt] + bc)$ (3)
3. Update gate: Update the old cell state, ct-1 information to new cell state ct. This is the new candidate values, scaled by to update each state value.
   $Ct = ft * ct-1 + it * ct$ (4)
4. Next output, it will be based on filtered version cell state. First sigmoid layer decides which parts of the cell state are going to be the output. Then put the cell state through tanh squish the values between -1 to +1 and multiply it by the output of the sigmoid function, so you can get only those parts you want.

GloVe: GloVe is an unsupervised learning algorithm for acquiring vector representations for words. GloVe model is trained on global collection of word-word co-occurrences statistics in a corpus, and the results show a linear substructure of words in vector space.

1. Word embeddings give a dense representation of words and their relative meanings.
2. They are an improvement over the sparse representations used in the simpler "bag of word" model representations.
3. Word embeddings can be learned from text data and reused among projects. They can also be learned as part of fitting a neural network onto text data.
4. Two different ways to apply Word Embeddings to a neural network
   - train embedding layer
   - use a pre-trained embedding (like GloVe).

IV. EXPERIMENTAL RESULT AND DISCUSSION

In Natural Language Processing (NLP), one of the widely

used tasks is text classification, whose goal is to automatically classify a text document into one or more predefined categories. It can also be called a supervised machine learning algorithm since the dataset is already labeled for training the classifier. The requirements to built a multi-headed model that can be capable of detecting different types of toxicity like threats, obscenity, insults, and identity-based hate. Data can belong to more than one category.

A. Dataset

The dataset is created by the Conversation AI team, a research initiative founded by Jigsaw and Google are working on tools to help improve online conversation. The current models still make errors, and they don't allow users to select different types of toxicity. The dataset is from Wikipedia's talk page edits. In this dataset, there are 160K comments and labeled with different categories some of the comments belong to more than one category.

Dataset Visualization : Data Visualization is a way of understanding the data by visual context.Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization.

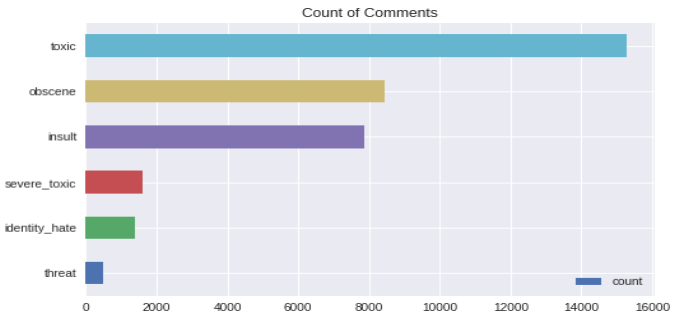

Fig 1. Bar chart of comments

This bar chart gives an idea about number of comments in different categories.

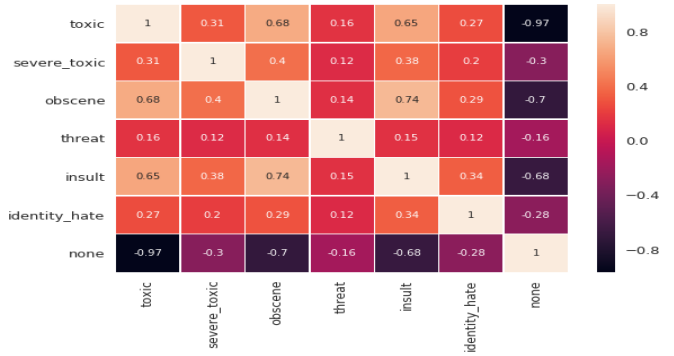

Fig 2. Correlation Matrix

The correlation matrix illuminates interesting relationships :

- "Toxic" comments are clearly correlated with both "obscene" and "insult" comments.
- Interestingly, "toxic" and "severe_toxic" are only weakly correlated.
- "Obscene" comments and "insult" comments are also highly correlated, which makes perfect sense.

B. Results

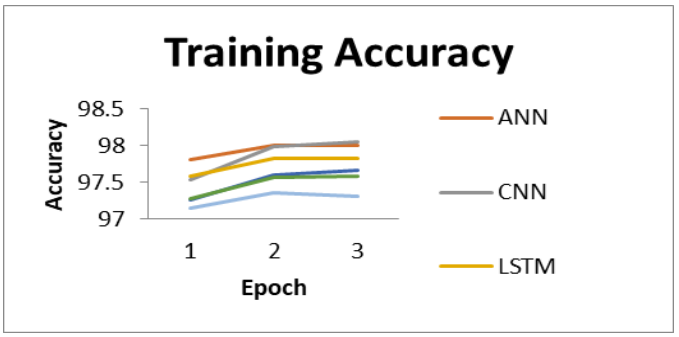

Fig 3: Accuracy on Training

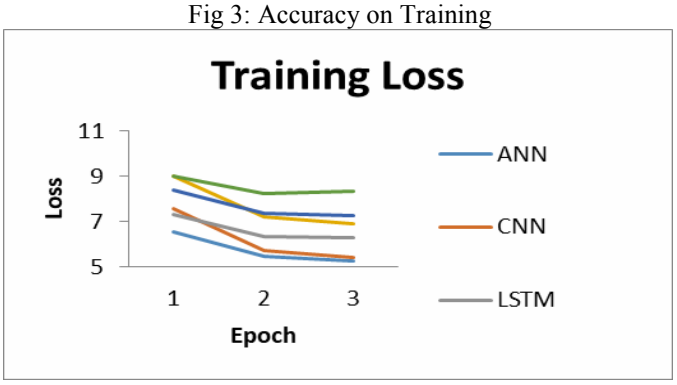

Fig 4: Loss on Training

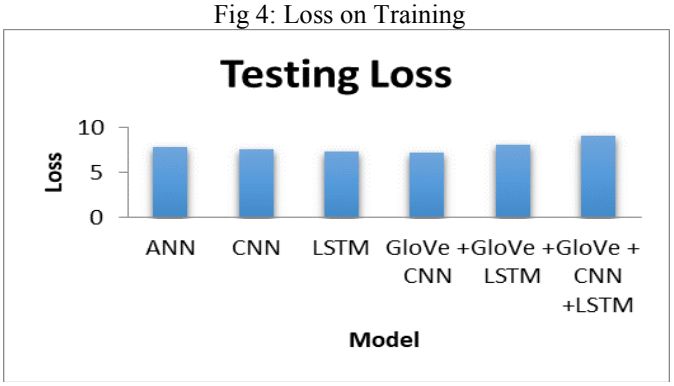

Fig 5: Accuracy on Testing

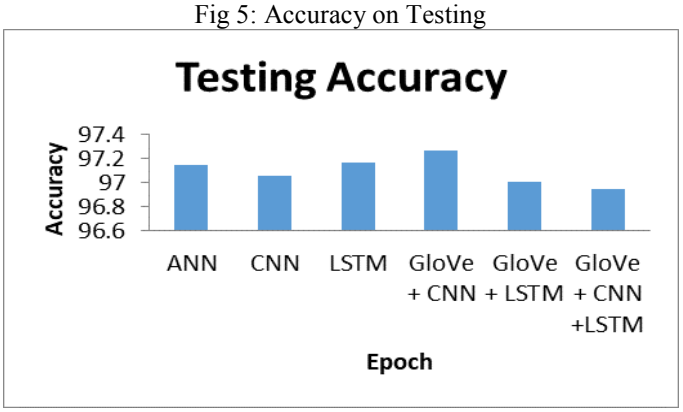

Fig 4: Accuracy & Loss on Testing

ANN: Gives 98% accuracy on training and minimum loss during loss. However the testing loss is high in terms of testing accuracy, third highest among all the models.

CNN: Training accuracy increases with each epoch and reach 97.8% and loss is 5.42%. However, in testing CNN gives

lower loss than ANN but Accuracy is less.
LSTM: Training Accuracy is lower than ANN and CNN, and the loss is higher than ANN and CNN. But in testing LSTM give better performance than previous models and its loss is also lower the previous models.
GloVe & CNN: Training accuracy is lower so the loss is on the higghher side than previous models. During testing the model performs better than all the models even its loss is similar to LSTM.
GloVe & LSTM: Training accuracy is on lower side so the loss is bit high. İn testing, it can be seen that it performs in terms of accuracy and loss.
GloVe & LSTM & CNN: Least accuracy and highest loss during training and testing. This model performs the worst than all other models.

TABLE I. COMPARISION TABLE OF ACCURACIES

MODEL	Training Accuracy	Validation Accuracy	Testing Accuracy
ANN	97.94	97.96	97.15
CNN	97.85	98.01	97.06
LSTM	97.75	98.07	97.19
GloVe & CNN	97.50	97.85	97.27
GloVe & LSTM	97.58	97.87	97.01
GloVe & CNN & LSTM	97.26	97.55	96.95

V. CONCLUSION

In this paper, a deep learning model is trained using various deep learning techniques to classify the comments in social meida netoworks in the following categories: toxic, severe toxic, obscene, threat, insult, and identity hate. Kaggle's toxic comment dataset is used for training. In conclusion, Glove & CNN performs the best and Glove & CNN & LSTM performs the worst in terms of training and testing, loss and accuracy. LSTM and ANN are performing the same followed by CNN and GloVe & LSTM. Recursive neural networks comprise a class of architecture that can operate on structured input. They have been prevoiusly successfully applied to model compositionality in natural language processing using parse tree based structural representation. It can be contructed by stacking multiple recursive layers. The results show thet deep RNNs outperforms the associated shallow counterpart the empoly the same number of parameters. Deep RNNs can be used for Abuse classification.

REFERENCES

[1] Pew Research Center: Online Harassment, http://www.pewinternet.org/2014/10/22/online-harassment/ , (2014)

[2] Yin, Z. Xue, 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, Madrid, Spain(2009)

[3] Sara Sood, Judd Antin, and Elizabeth Churchill : Using crowdsourcing to improve profanity detection. In 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. In Proceedings of the 25th International Conference on World Wide Web (WWW '16)., \emph{}, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 145-153. DOI: https://doi.org/10.1145/2872427.2883062, (2016)

[5] Ellery Wulczyn, Nithum Thain, and Lucas Dixon: Ex Machina: Personal Attacks Seen at Scale. In Proceedings of the 26th International Conference on World Wide Web (WWW '17), International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1391-1399. DOI: https://doi.org/10.1145/3038912.3052591 (2017)

[6] Wulczyn, Ellery; Thain, Nithum; Dixon, Lucas: Wikipedia Detox. figshare. doi.org/10.6084/m9.figshare.4054689 (2016)

[7] Yoon Kim: Convolutional neural networks for sentence classification. In EMNLP. Association for Computational Linguistics, 1746—1751 (2014)

[8] Yoon Kim, "Convolution Neural networks for sentences classification", In EMNLP, Association for computational Linguistics, 1746-1751(2014)

[9] Amir H. Razavi, Diana Inkpen, Sasha Uritsky , Stan Matwin: Offensive Language Detection Using Multi-level Classification. 23rd Canadian conference on Advances in Artificial Intelligence pages 16-27.(2010)

[10] Ellery Wulczyn, Nithum Thain, Lucas Dixon: Ex-Machina: Personal Attacks seen at scale. Published in WWW2017. DOI: 10.1145/3038912.3052591

[11] Ji Ho Park, Pascale Fung: One step and two step classification for abusive language detection on twitter. http://arxiv.org/abs/1706.01206

[12] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yeshar Mehdad, Yi Chang: Abusive Language Detection in online user content. 25th International conference on World Wide Web, pages:145-153