# Self-Attention for Cyberbullying Detection

Ankit Pradhan, Venu Madhav Yatam, and **Padmalochan Bera**

IIT Bhubaneswar

June 17, 2020

# Introduction

- Cyberbullying has been a very important issue in the age of internet and the digital revolution. It has been defined as bullying through the Internet or using social media, messaging and gaming platforms or more recently, mobile phones, to repeatedly embarrass or hurt people.
- With psychological impact as adverse as offline bullying, victims have potential chances of falling into prolonged depression, isolate themselves from the society or even entertain suicidal thoughts.
- Adolescents are at an higher risk of being subjected to cyberbullying as pointed out many recent news articles.

# Introduction (contd...)

Cyberbullying detection becomes essential for ensuring social well-being.

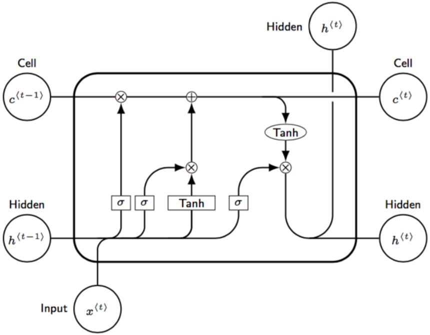
The factors that make cyberbullying detection challenging are :

- Subjective nature
- Association with diverse topics such as religion, gender, colour, etc.
- Vocabulary and comprehension of words vary over multiple social media platforms.

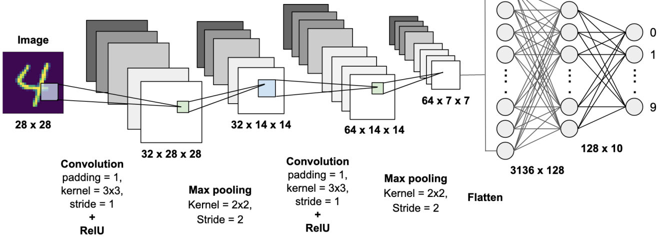
# Related work

- S. Hochreiter, and J. Schmidhuber, Long short-term memory. Neural Computation, 1997, 9(8), 1735–1780.
- Y. Kim, “Convolutional Neural Networks for Sentence Classification,” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), pp. 1746–1751, 2014.
- S. Amir, B. C. Wallace, H. Lyu, P. Carvalho, and M. J. Silva, “Modelling Context with User Embeddings for Sarcasm Detection in Social Media,” Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL), pp. 167–177, 2016.
- A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, “Very Deep Convolutional Networks for Natural Language Processing,” In Proceedings of EACL, volume 1, pages 1107–1116.

# LSTM & CNN architectures



(a) Structure of LSTM



(b) Generic CNN architecture

Figure 1: Architectures of LSTM (Long Short Term Memory) & CNN (Convolutional Neural Networks)

# C-LSTM architecture

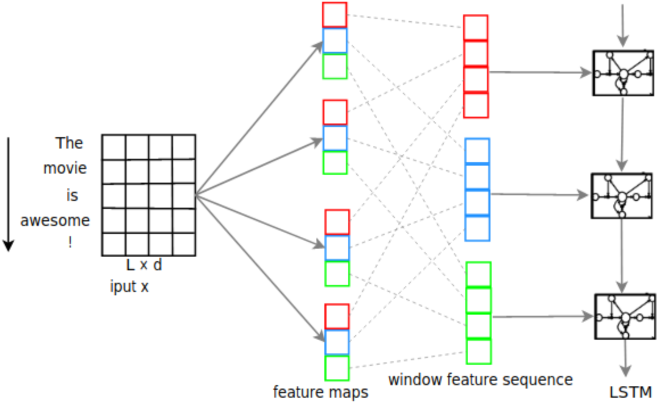


Figure 2: Hybrid C-LSTM by Zhou et al. [3]

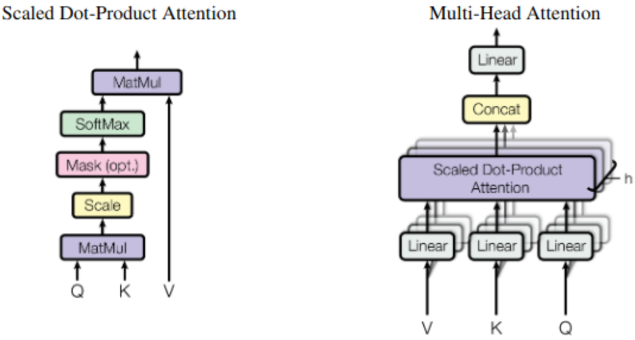
# Attention mechansim

The attention mechanism in deep learning (DL) is a method to allocate attention values to various components of a system quantitatively.

An attention function maps a query and a tuple of key-value pairs to an output. Here, the weights assigned to each value is computed by a helper function with a corresponding key input. Finally, a weighted sum of the input values is given as the output. In general, there are two types of attentions:

- General Attention: This kind of attention is between several input elements in a layer of a neural network. These attention mechanisms differ based on the alignment score used.
- Self Attention: Self-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the same sequence. This is also known as intra-attention.

# Scaled Dot-Product & Multi-Head Attentions



Scaled Dot-Product Attention

Multi-Head Attention

Figure 3: Scaled Dot-Product and Multi-Head Attentions proposed in [1]

# Model architecture

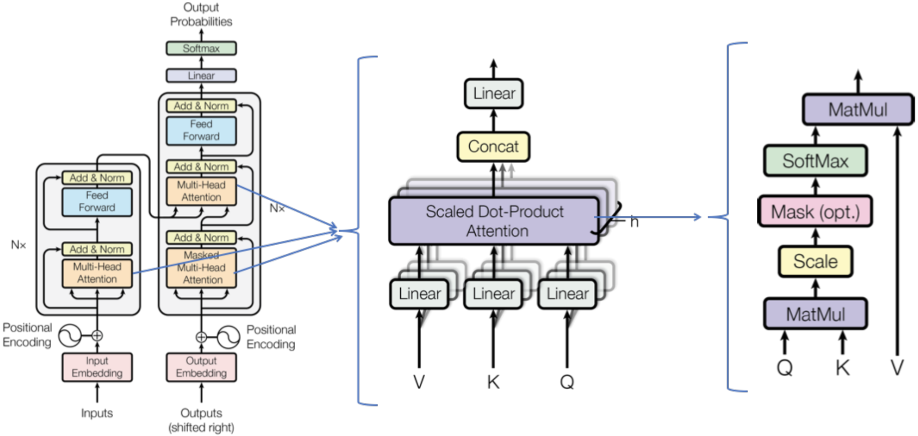


Figure 4: Self-attention with Transformer architecture proposed in [1]

# Experimentation

This work uses three open source datasets : Wikipedia dataset, Formspring dataset, and Twitter dataset.

- **Wikipedia dataset**: This dataset contains around 100k labeled discussion comments from Wikipedia's talk pages.
- **Formspring dataset**: Formspring is a social network based on questions and their corresponding answers. It was launched in 2009 and has been a standard dataset for cyberbullying works.
- **Twitter dataset**: The dataset is created by performing a manual initial search of common illicit words and terms used corresponding to sexual, gender, religious and ethnic minorities and discrimination. This dataset contains 16K annotated tweets.

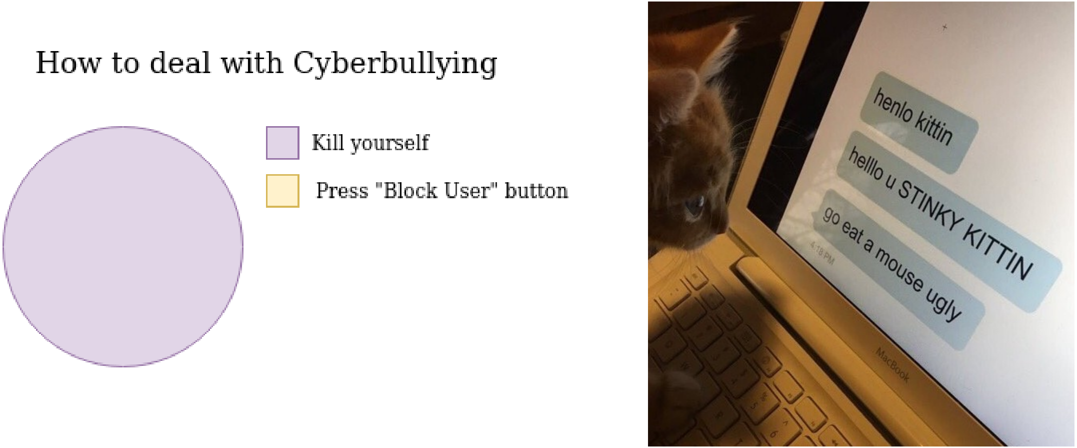
# Results & Conclusion

| Models | Datasets | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Wikipedia | | | Formspring | | | Twitter | | |
| | Precision | Recall | F1-score | Precision | Recall | F1-score | Precision | Recall | F1-score |
| SVM | 0.747 | 0.686 | 0.723 | 0.714 | 0.718 | 0.716 | 0.814 | 0.808 | 0.811 |
| Logistic Regression | 0.645 | 0.628 | 0.634 | 0.709 | 0.714 | 0.711 | 0.808 | 0.817 | 0.812 |
| CNN by Kim et al. (2014) [2] | 0.793 | 0.786 | 0.786 | 0.819 | 0.816 | 0.817 | 0.813 | 0.806 | 0.808 |
| C-LSTM by Zhou et al. (2015) [3] | 0.748 | 0.762 | 0.754 | 0.798 | 0.799 | 0.799 | 0.845 | 0.842 | 0.843 |
| BLSTM using attention (2018) [4] | 0.810 | 0.670 | 0.740 | 0.560 | 0.490 | 0.510 | 0.740 | 0.760 | 0.750 |
| Self-Attention | 0.842 | 0.838 | 0.841 | 0.894 | 0.915 | 0.905 | 0.897 | 0.885 | 0.891 |

Table 1: Precision, Recall and F1-scores of our system for the three datasets vs other models

The Self-Attention model outperforms the BLSTM model with attention. The Precision, Recall and F1- scores for Twitter dataset obtained are 89.7%, 88.5% and 89.1% respectively which are around a 5-6% increase from the best model. In case of Formspring dataset, these values are 89.4%, 91.5% and 90.5% respectively which are around 7-8% more than the best model. Wikipedia dataset also exhibits similar performance with corresponding values of 84.2%, 83.8% and 84.1%.

# Future work



henlo kittin

helllo u STINKY KITTIN

go eat a mouse ugly

Figure 5: Cyberbullying meme in the form of pie-charts & as text embedded in image

# Bibliography

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N Gomez,L. Kaiser, and I. Polosukhin,
Attention Is All You Need
In *Advances in Neural Information Processing Systems 2017*, pages 6000–6010

Y. Kim
Convolutional Neural Networks for Sentence Classification
In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp. 1746–1751, 2014.

C. Zhou, C. Sun, Z. Liu and F. C. M. Lau.
A C-LSTM Neural Networkfor Text Classification
In *Arxiv*, CoRR, abs/1511.08630, Nov 2015.

S. Agrawal and A. Awekar.
Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms
In *Pasi G., Piwowarski B., Azzopardi L., Hanbury A. (eds) Advances in Information Retrieval, ECIR 2018*, Lecture Notes in Computer Science, vol 10772, Springer, 2018.