Odium Revelio! Detecting Subtle Hate Speech in Online Conversations

Sakshi Agarwal[+] (sakshia1@uci.edu), **Sandya Mannarswamy*** (sandyasm@gmail.com)

[+] University of California, Irvine, * Independent Researcher

Introduction

Misuse of social media has the power to induce trigger hatred, abuse and toxicity.

Example –
Comment text *: Ever heard of the republican icon XYZ?*
Reference *: XYZ accused of sending sexual emails to young boys.*

Do the automated approaches detect the subtle expression of hate speech?

No!
While neural network models have been proposed for hate speech classification, they have not modeled this problem.

Objectives

- Model background knowledge.
- Enrich an existing dataset with additional background information.
- Develop a neural network based approach and evaluate on dataset.

Dataset?

Fox News User Comments corpus[1]

- **1528 comments**
- **678 unique users**
- **10 News articles**
- **435 hateful comments**

Background information?

- Wikipedia , UrbanDictionary (manual or query matching)
- Article Summary, Previous comments

What is Inter-attention?

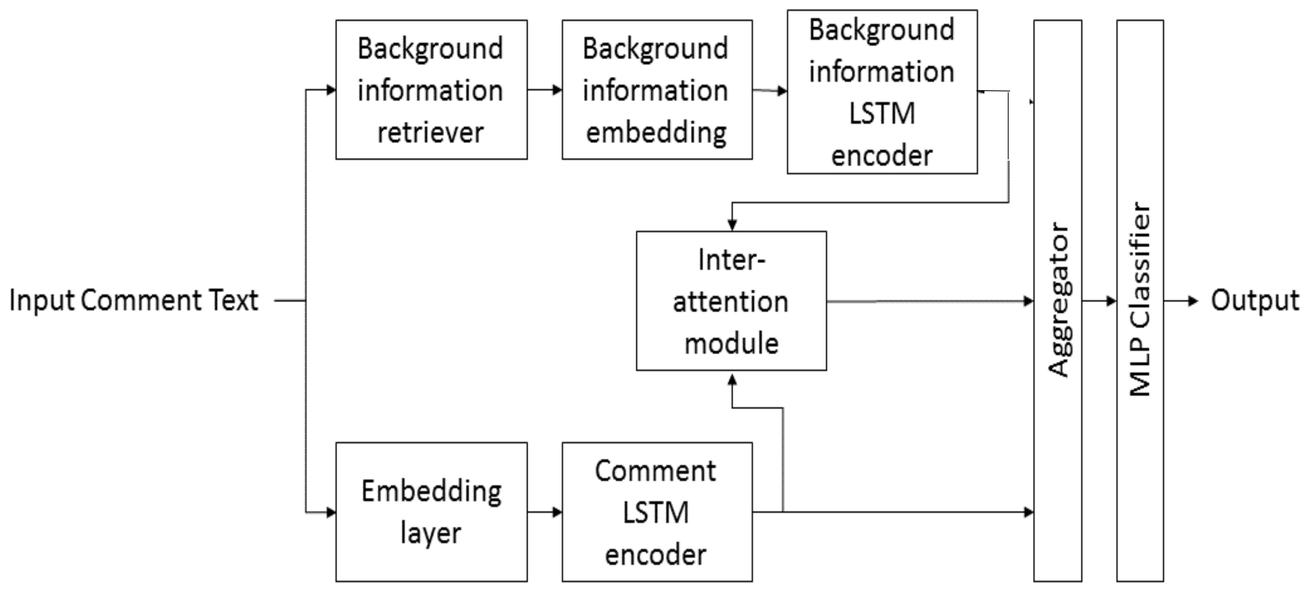
Attention weights of an input text encodings are learnt from an encoded representation of a related text.

Equations for generating the cross-text interactions :

(i) $M_1 = tanh(W_1 Y_c + W_2 Y_B)$
(ii) $\alpha = softmax(W^T M_1)$
(I) $O_1 = \alpha Y_c$

Methodology

- Encode comment text (T), background information into neural sentence embedding Y_C, Y_B respectively.
- Capture the cross-text interaction (O_1) between Y_C and Y_B.
- Feed O1 to a standard Multi-layer Perceptron (MLP) classifier.



Functional Components of our Approach

Experiment Setup

- TensorFlow
- Pre-trained word embedding
- Cross-entropy loss
- Adam Optimizer
- Length of article title – 60
- Length of summary – 150.
- Length of T = 150
- Length of external information - 184.
- Softmax layer followed by fully connected MLP classifier.

Results

- Two baseline classifier -
  SVM
  Gradient boosting
  Bag of words model
  Lexicon : hatebase.org
- Comparisons between intra-attention and inter-attention.

Method	Accuracy	Precision	Recall	F-Score
SVM	0.68	0.61	0.68	0.60
Gradient-Boost	0.70	0.67	0.70	0.59
No external	0.75	0.65	0.42	0.51
Intra-attention	0.81	0.73	0.63	0.68
Inter-attention	**0.85**	**0.84**	**0.85**	**0.84**

Future Work

- Handle “sound alike” hateful comments like “just like Milk, this ship will be full of sea-men”.
- Incorporate image content on posts as additional information.

Acknowledgements

I thank Conduent Labs, Bangalore for giving me the opportunity to work on this project. I am very thankful to my mentor, Sandya who guided me throughout this research.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. CoRR abs/1409.0473.
Burnap, P., and Williams, M. L. 2015. Cyber hate speech on twitter : An application of machine classification and statistical modeling for policy and decision making
Gao, L., and Huang, R. 2017. Detecting online hate speech using context aware models. CoRR abs/1710.07395
Davidson, T.; Warmsley, D.; Macy, M. W.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. CoRR abs/1703.04009

1 https://github.com/sjtuprog/fox-news-comments