Myanmar Hate Speech Detection

Phyo Thu Htet

University of Technology (Yatanarpon Cyber City)

phyothuhtet@studentpartner.com

February 23, 2020

Abstract

As an adverse point of social media, these tools and platform can be used as outlets to spread hate speech. A stimulus reinforced with 'AI' technology for detecting Myanmar hate speech would be one of the crucial mechanisms to address the escalation of virulent hate speech which has negative butterfly effect on the country of Myanmar.

Key Words: Hate Speech, Neural Network, Sentiment Analysis, NLP, Myanmar

Problems

- Vulnerable victims around the world suffer the hate speech attacks which is directly proportional to the rise of social media and other means of online communication tools.
- However, cases like persecutions can be involved in posting hate speech to stimulate or provoke negative impacts delivering harassment, racism, denigration,oppression and also marginalization or social exclusion.
- And Myanmar text sentences are not prima facie ones to distinguish hate speech or not - even heading to dilemmas.

Objectives

1

To study which AI models and processes can deliver the better results in detecting Myanmar Hate Speech

2

To remove or trace Myanmar hate speech sentences by auto-detecting as the first step considering from the view point of technical dimension

3

To take a part in helping to achieve auspicious guidances to end hate speech

Definition

Facebook's Hate speech Definition

"We define hate speech as a direct attack on people based on what we call protected characteristics—race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability."

Sample Data

HateSpeech	NormalSpeech
မဲလိုက်တာ အသားအရေက အဘွားကြီးကို မဲပေးဖို့ ခွေးတွေ စောင့်ကြတယ် တစ်ခွန်းပဲပြောမယ် မသာမ	အာ့အရုပ်လေးလိုချင်တယ် သွားပီးfollowထားအုံးမှပါ စာတေရတော့မာပဲ

Table: Hatespeech vs Normalspeech

Methodology

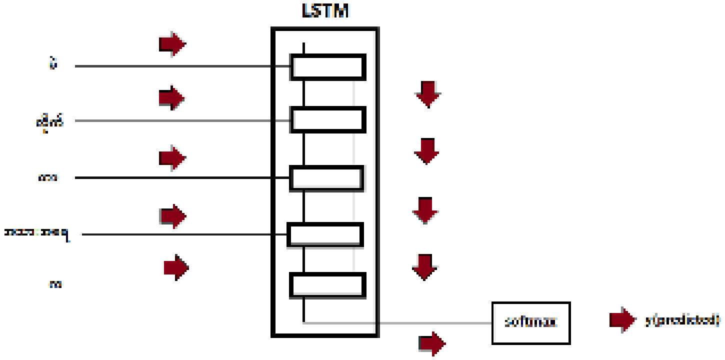

Fig: Model Architecture

Con't

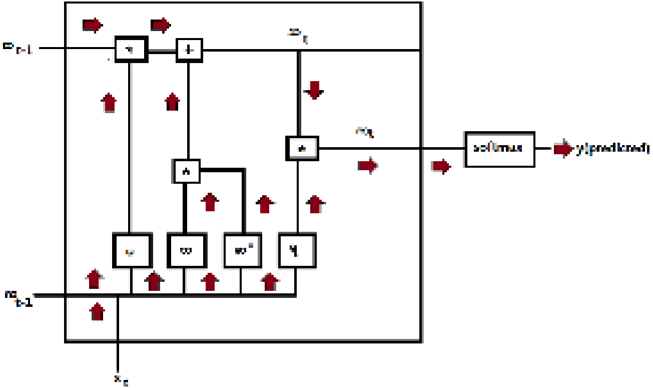

Fig: More detail architecture of lstm

Con't

Theorem (Equations)

$${}^{*}_{t} = \mathrm{tanh}\left(W, x_t + b\right)$$

$$\text{ဖ} = \sigma\left(\mathsf{W}[{}_{t-1}, x_t] + b\right)$$

$$\text{ထ} = \sigma\left(\mathsf{W}[{}_{t-1}, x_t] + b\right)$$

$$\text{ရ} = \sigma\left(\mathsf{W}[{}_{t-1}, x_t] + b\right)$$

$${}_t = * \, {}^{*}_{t} + * \, {}_{t-1}$$

$${}_t = * \, tanh\left({}_t\right)$$

Con't

Example (Exaplanation of variables)

ဆ = memory cell

ဖ = forget gate

သ = update gate

ရ = output gate

က = activation

References

Facebook
Hate Speech Definition

Andrew Ng
Sequence Model

Hao-Ren Yao,Eugene Yang,Katina Russell,Nazli Goharian,Ophir Frieder
Hate Speech Detection: Challenge and Solution

Gracias

ကျေးဇူးပါ