# Deep Encrypted Text Categorization

Vinayakumar R[1], K.P Soman[1] and Prabaharan Poornachandran[2]

[1]Centre for Computational Engineering and Networking (CEN), Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham,

Amrita University, India.

[2]Center for Cyber Security Systems and Networks, Amrita School of Engineering, Amritapuri, Amrita Vishwa Vidyapeetham,

Amrita University, India.

# Outline

- Introduction
- Methodology
- Description of the data set and Results
- Summary
- Future Work
- References

# Introduction

- Text categorization focuses on classifying text to its categories and it has roots in many natural language processing (NLP) applications, mainly in content security.
- Content security is an approach used by network administrator to safeguard internet security from malicious attacks with the text available in online.

# Methodology

- Text categorization has been existing as a difficult task mainly due to the traditional machine learning approaches relies on bag-of-words (BoW) model or bag-of-n-gram vectors, where unigrams, bigrams, n-grams, punctuation, stop words, emoticons, negation words, lexicons, elongated words and other delineated patterns are considered as features.
- To alleviate, we have used word embedding with deep learning specifically recurrent neural network and long short-term memory.

# Contd.

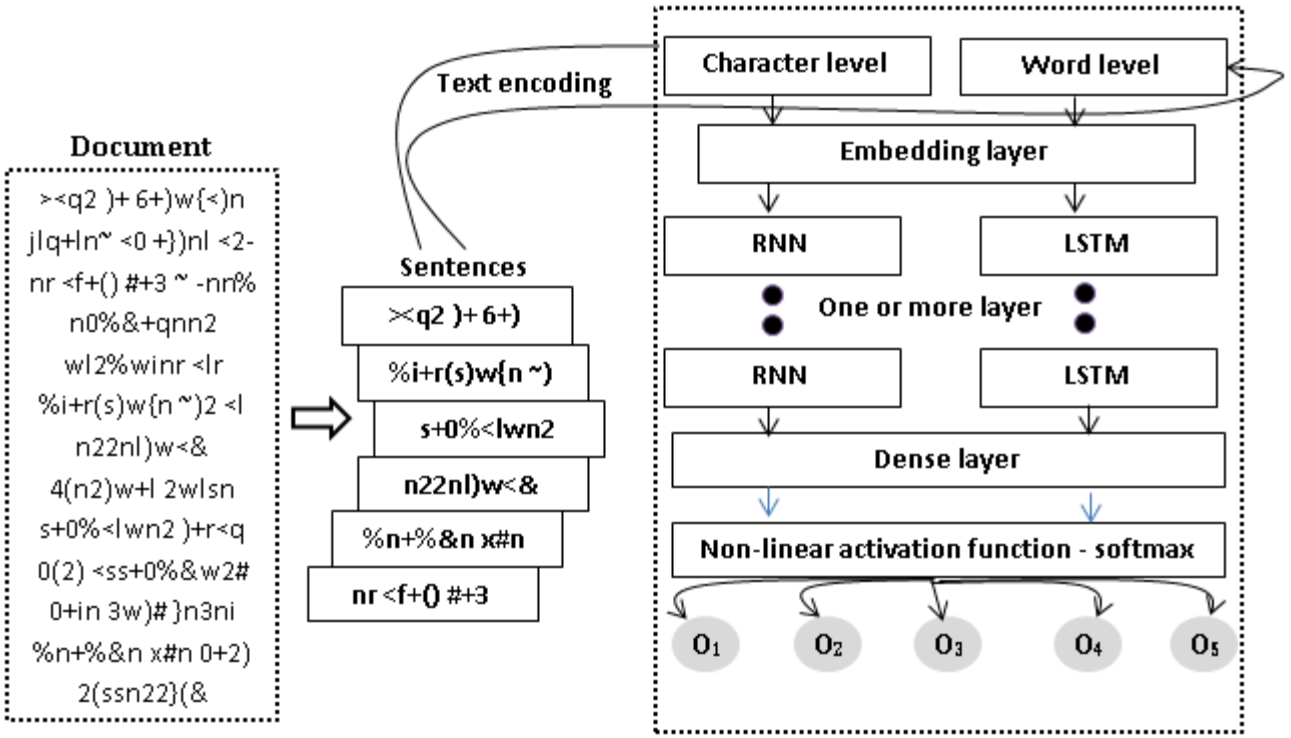


Figure 1. proposed deep learning architecture for encrypted text categorization

# Description of the data set and Results

The dataset is sourced from 6 online news media: The New Zealand Herald [1], Reuters [2], The Times [3] , Yahoo News [4], BBC [5] and The Press [6]. Business, entertainment, sport, technology, and travel are the selected five news categories.

Table 1. Description of Data set

| Topic | Number of News |
|---|---|
| Business | 361 |
| Entertainment | 343 |
| Sport | 363 |
| Technology | 356 |
| Travel | 362 |

# Contd.

| Architecture | Input Type | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|---|
| RNN 1 layer | Character | 0.964 | 0.999 | 0.956 | 0.977 |
| LSTM 1 layer | Character | 0.987 | 0.959 | 0.947 | 0.953 |
| RNN 2 layer | Character | 0.873 | 0.863 | 0.966 | 0.912 |
| LSTM 2 layer | Character | 0.891 | 0.952 | 0.220 | 0.357 |
| RNN 1 layer | Word | 0.940 | 0.880 | 0.995 | 0.934 |
| LSTM 1 layer | Word | 0.973 | 0.944 | 0.996 | 0.969 |
| RNN 2 layer | Word | 0.858 | 0.837 | 0.983 | 0.904 |
| LSTM 2 layer | Word | 0.914 | 0.838 | 0.992 | 0.909 |

Table 2. 5-fold cross-validation results of RNN and LSTM networks

# Contd.

| Overall accuracy | Class wise accuracy | Precision | Recall | Specificity |
|---|---|---|---|---|
| 0.43 | 0.772 | 0.442 | 0.432 | 0.858 |

Table 3. Summary of test results using 3 layer stacked LSTM network with 32 memory blocks

# Summary

- The proposed deep learning architecture for encrypted text categorization avoids the feature engineering method, thereby itself serves as a robust in handling drifting of encrypted texts in the scenario of adversarial machine learning setting.
- LSTM network performed well in comparison to the RNN.

# Future Work

- we are lack behind in showing the experimental results for encrypted text categorization with more complex architecture. The reported results can be further enhanced by using more complex architecture by using an advanced hardware and training in a distributed environment.

# References

[1] www.nzherald.co.nz

[2] www.reuters.com

[3] www.timesonline.co.uk

[4] news.yahoo.com

[5] www.bbc.co.uk

[6] www.stuff.co.nz