CIT Kokrajhar Team: LSTM based Deep RNN Architecture for Hate Speech and Offensive Content (HASOC) Identification in Indo-European Languages

Baidya Nath Saha[1] and Apurbalal Senapati[2]

[1] Concordia University of Edmonton, Edmonton AB T5B 4E4, Canada
baidya.saha@concordia.ab.ca
[2] Central Institute of Technology, Kokrajhar BTAD, Assam 783370, India
a.senapati@cit.ac.in

Abstract. Recently, automated hate speech and offensive content identification has received significant attention due to rapid propagation of cyberbullying which undermines objective discussions in social media and adversely affects the outcome of the online social democratic processes. A special type of Recurrent Neural Network (RNN) based deep learning approach called Long Short Term Memory (LSTM) is implemented for automatic hate speech and offensvie content identification. Separating offensive content is quite challenging because the abusive language is quite subjective in nature and highly context dependent. This paper[3] offers language-agnostic solution in three Indo-European languages (English, German, and Hindi) since no pre-trained word embedding is used. Experimental results offer very attractive insights.

Keywords: Hate Speech Dtection · Offensive Content Identification · Long Short Term Memory (LSTM) · Recurrent Neural Network (RNN).

1 Introduction

Social media communications induce strong impact on public opinion and some social platforms possess enough social capital to influence the outcome of democratic processes [8]. However, hate speech and offensive language such as insulting, hurtful, derogatory and objectionable obscene content seriously affect the dynamics and usefulness of online social communities. Public communication potentially realizes substantive rational critical discourse [3]. Hate and offensive content widely circulated in social media rejects or subverts the status quo of objective discussions, leads to the polarization, stigmatization and radicalization of public debates, and poses a potential threat to democratic society. However,

open societies need to adopt a means to restrict hate speech and offensive content by addressing specific issues about intimidation or incitement, without prohibiting the free exercise, and thereof abridging the freedoom of speech and enforcing general social regulation. As a consequence, many platforms of social media websites follow a standard measure by manually reviewing hefty online contents to identify and delete offensive materials which is a cumbersome process and not sustainbale in reality. This leads to a pressing need for developing scalable, automated methods for HAte Speech and Offensive Content (HASOC) identification and has attracted significant research using semantic content analysis based on Natural Language Processing (NLP) and Machine Learning (ML).

Automating its' detection and then adpoting proper countermeasures could reduce the propagation of HASOC significantly. However, it encounters severe challenges such as disagreements in defining hate speech depending on it's subjectivity and context-dependent characteristics which make difficult to separate hate speech from the remaining texts [5]. From different cultural perspectives, using abusive words while expressing opinion are not always regarded as insulting or inciting hatred which indicates that some content can be considered hate speech to some and not to others, based on their respective definitions. On the other hand, hate speech does not always contain offensive words while offensive language does not always express hate. In addition, nuance and subtleties in language provide further challenges in automatic hate speech identification, however, it also depends on the definition.

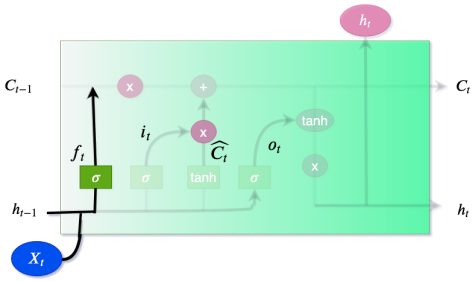
This research portrays the HASOC identification problem as sentence classification problem similar to sentiment analysis which is an active research area over the last few years [7]. Several researches have been founded on HASOC identification in English. However, there are limited works found in other Indo-European languages such as German and Hindi. Our approach employs a neural network solution based on Long-Short-Term-Memory (LSTM). The intended methodology does not use any pre-trained word embedding which provides language-agnostic solution.

2 Data description

Dataset has been created by FIRE 2019 organizers for the HASOC identification shared task from the Twitter and Facebook in three Indo-European (German [9] English and code-mixed hindi) and distributed in TSV (Tab Separated Value) format [6]. Training data corpus for English, German, and Hindi consist of 5853, 3820, and 4666 sentences respectively. Test data corpus for English, German, and Hindi contain 1153, 850, and 1318 sentences respectively. There are three tasks available in the dataset [11], [4], [10]. Task 1 focuses on HASOC identification offered for English, German, Hindi. Task 1 is coarse-grained binary classification where tweets data are classified into two class, namely: Hate and OFfensive (HOF) and Non- hate and offensive (NOT). HOF post contains hate, offensive, and profane content. A post is annotated as HOF if it contains any form of non-acceptable language such as hate speech, aggression, and profanity.

NOT post does not contain any hate speech, and offensive content. Task 2 is a fine-grained classification. For task 2, hate-speech and offensive posts from the task 1 are further classified into three categories: HATE speech (HATE): posts under this class contain hate speech content, OFFeNive (OFFN): posts under this class contain offensive content, and PRoFaNe (PRFN): these posts contain profane words. Hate speech describes negative attributes or deficiencies to groups of individuals because they are members of a group (e.g. all poor people are stupid). These are hateful comment toward groups because of race, political opinion, sexual orientation, gender, social status, health condition or similar. Then degrading, dehumanizing, insulting an individual, threatening with violent acts are categorized into offensive category. Profanity is the unacceptable language in the absence of insults and abuse. This typically concerns the usage of swearwords (Scheiße, Fuck etc.) and cursing (Zur Hölle! Verdammt! etc.) are categorized into PRFN category. Most posts belong to OTHER category, some are HATE and the other two categories are less frequent. Dubious cases which are difficult to decide even for humans, are left out. Task 3 examines the type of offense. Only posts labeled as HOF in task 1 are included in task 3. The two categories in task3 include: (a) Targeted Insult (TIN): posts containing an insult/threat to an individual, group, or others, and (b) Untargeted (UNT): posts containing nontargeted profanity and swearing. Posts with general profanity are not targeted, but they contain non-acceptable language.

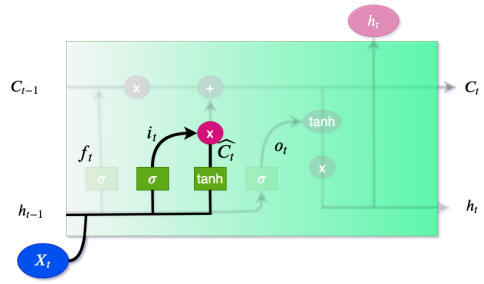
3 Methodology



$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

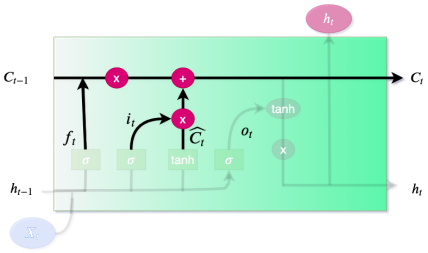
Fig. 1. Sigmoid layer. Inspired from [1]

We developed a special type of Recurrent Neural Network (RNN) [2] based deep learning approach called Long Short Term Memory (LSTM) to detect hate speech and identify offensive content in three Indo-European languages (Hindi, English, and German). RNN is a family of neural networks used for processing of sequential data. Traditional RNN can learn the model of complex spatio-temporal dynamics by mapping the input sequence to a sequence of hidden states, and the exit of the hidden states to the output layer.



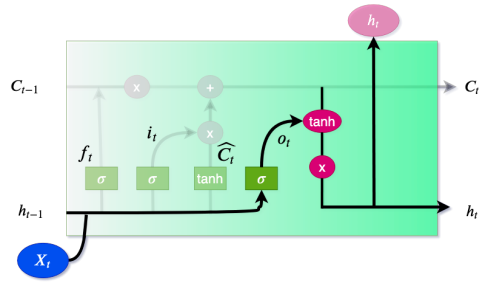
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$\widehat{C_t} = tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

Fig. 2. Sigmoid and Tanh layers. Inspired from [1]



$$C_t = f_t * C_{t-1} + i_t * \widehat{C_t}$$

Fig. 3. New cell state in LSTM. Inspired from [1]



$$o_t = \sigma(W_o[h_{t-1}, h_t] + b_o)$$
$$h_t = o_t * tanh(C_t)$$

Fig. 4. Output layer. Inspired from [1]

The first step in LSTM network is to decide the information we are going to keep so that it continues its path throughout the cell state. This decision is achieved using a sigmoid layer called the layer of the forget gate. This layer can be seen in Fig. 1 where the information is discriminated against when using a sigmoid function. After this, the information that survived this process enters a new layer where it is processed again by a sigmoid function and then reconfigured in

another layer with a hyperbolic tangent function in order to create a vector of new candidate values as observed in the Fig. 2.

Then the old cell state, C_{t-1} is updated into the new cell state C_t which is demonstrated in Fig. 3. Finally, this new vector is filtered again using a combination of sigmoid activation function and tangent hyperbolic as can be seen in Fig. 4. The following notations are used for describing the LSTM netwotk:

- $x_t \in \mathbb{R}^d$ is the input vector of LSTM unit
- $f_t \in \mathbb{R}^h$ activation vector (forget gate)
- $i_t \in \mathbb{R}^h$ input and update of the gate vector
- $o_t \in \mathbb{R}^h$ exit of the activation
- $h_t \in \mathbb{R}^h$ hidden state vectors (exit vector of the LSTM unit)
- $\widehat{C_t} \in \mathbb{R}^h$ New candidate vectors for cell status
- W y b Weight matrices and bias vectors that need to be learned during network training

It is to be noted that Figures 1, 2, 3, and 4 are inspired from the online tutorial "Understanding LSTM Networks" [1].

4 Experimental results and discussions

In our experiment, we exploited the following architecture:

word embedding → LSTM with hidden layer → Dense layer with a neuron → sigmoid activation function.

Vectors length 128 for embeddings layer, 128 neurons in each hidden layer, batch size 60, 10 number of epochs and a dropout of 20% were chosen for this experiment. In order to establish the convergence of the network, binary and categorical cross entropy type error function were used for two- and multi-class classification respectively. ADAM optimizer was used for all the tasks classification. We used the default parameters of Keras for ADAM optimizer.

Results of LSTM based deep RNN architecture for task 1, task 2, and task 3 associated with HASOC identification are mentioned in Table 1, Table 2, and Table 3 respectively. Binary cross-entropy loss function is used for binary sentence classification: hate and offensive (HOF) and non hate-offensive classification (NOT) associated with task 1. Categorical cross-entropy loss function is used for multiclass sentence classification associated with task 2: hate speech (HATE), offensice (OFFN), profane (PRFN), and NOT and task 3: targeted insult (TIN), untargeted insult (UNT), and NOT. Table 1, Table 2, and Table 3 demonstrates the performance of LSTM based RNN claasifier in terms of weighted average of accuracy, precision, recall and F-measure. Results demonstrate that LSTM based RNN classifier performs better in German than English and Hindi dataset.

Table 1. Classification results for task 1.

Language	Accuracy	Precision	Recall	F-measure
Hindi	0.53	0.54	0.53	0.53
English	0.52	0.60	0.52	0.55
German	0.81	0.73	0.81	0.76

Table 2. Classification results for task 2.

Language	Accuracy	Precision	Recall	F-measure
Hindi	0.41	0.41	0.41	0.41
English	0.52	0.57	0.52	0.55
German	0.79	0.72	0.79	0.75

5 Conclusion

Long Short Term Memory (LSTM) based Recurrent Neural Network (RNN) is implemented for automatic HASOC identification in three Indo-European languages: English, German, and Hindi. The potential challenge of HASOC detection comes from its subjectivity and context-dependent characteristic. Proposed methodology does not utilize any pre-trained model which leads to language-neutral solution as well as there is no unwieldy feature engineering required for the proposed model. In future, we would like to exploit other deep learning models for HASOC identification.

References

1. Understanding lstm networks. https://colah.github.io/posts/2015-08-Understanding-LSTMs/, accessed:September, 2019
2. Graves, A.: Supervised Sequence Labelling with Recurrent Neural Networks (2011)
3. Habermas, J.: The Theory of Communicative Action, Volume 1: Reason and the Rationalization of Society. Boston: Beacon Press (1984)
4. Kumar, R., Ojha, A.K., Malmasi, S., Zampieri, M.: Benchmarking aggression identification in social media. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018) (aug 2018)

Table 3. Classification results for task 3.

Language	Accuracy	Precision	Recall	F-measure
Hindi	0.50	0.53	0.50	0.51
English	0.54	0.59	0.54	0.56

5. Lee, Y., Yoon, S., Jung, K.: Comparative studies of detecting abusive language on twitter. CoRR **abs/1808.10245** (2018), http://arxiv.org/abs/1808.10245
6. Modha, S., Mandl, T., Majumder, P., Patel, D.: Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (2019)
7. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1-2), 1–135 (Jan 2008). https://doi.org/10.1561/1500000011, http://dx.doi.org/10.1561/1500000011
8. Pitsilis, G., Ramampiaro, H., Langseth, H.: Effective hate-speech detection in twitter data using recurrent neural networks. Applied Intelligence p. in press. (07 2018). https://doi.org/10.1007/s10489-018-1242-y
9. Wiegand, M., Siegel, M., Ruppenhofer, J.: Overview of the germeval 2018 shared task on the identification of offensive language. In: Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018) (09 2018)
10. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Predicting the type and target of offensive posts in social media. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 1415–1420. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1144, https://www.aclweb.org/anthology/N19-1144
11. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 75–86 (2019)