# Sentiment Analysis with Explainable AI Method Layer-based Relevance Propagation (LRP)

Mahmudul Haque
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
mahmudul.haque@g.bracu.ac.bd

Sahiba Tasneem
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
sahiba.tasneem@g.bracu.ac.bd

Humayra Ferdousi
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
humayra.ferdoushi@g.bracu.ac.bd

***Abstract* - In recent year text or sentiment analysis has become essential research for many online, social media market research to customer services. To systematically identify the exact sentiment, we propose to use a popular XAI technique called Layer-wise Relevance Propagation (LRP). We extended this explainable AI's usage towards a word-based bi-directional LSTM model with a five-class sentiment prediction task. We propose to use the Google API translator to evaluate the result in LRP relevance in the language of Bangla.**

***Keywords- Sentiment Analysis, Opinion Mining, XAI LRP***

## I. Introduction

Sentiment analysis, known as emotion AI, refers to natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and personal information. The basic task in sentiment analysis is classifying a given text's polarity at the document, sentence, or a single text whether the expressed opinion is positive, negative, or neutral. Advanced sentiment classification looks, for instance, at emotional states such as angry, happy, or sad.

Recently, a trend emerged for tackling these challenges via machine learning models and recurrent neural networks as observed e.g. Explaining Recurrent Neural Network Predictions in Sentiment Analysis (Arras, Montavon, Muller, & Samek, 2019)

The technical challenge of expression AI decisions are known as the interpretability problem. To get the correct text polarity check, In our paper, we used the most popular XAI LRP. This technique hides the undesirable attributes by allowing a tradeoff between the interpretability and completeness of an explanation.

Neural network interpretability, established as an important and active research area. Among many approaches to explaining the recurrent neural network predictions, the Layer-wise Relevance Propagation (LRP) framework has been proven successful at providing a meaningful intuition and measurable quantities describing a network's feature processing and decision making. (Kohlbrenner, et al., 2020)

In this paper, we implemented the LRP rule, which handles multiplicative interactions in the Long-Short Term Memory (LSTM) model, a suitable model for long-range expressions in a sentence such as those occurring in sentiment analysis. The bi-directional LSTM trained on five different level sentiment prediction job. These levels allowed us to attribute different sentiment in individual texts, compared to other explanation methods. (Arras, Montavon, Muller, & Samek, 2019)

## II. Related Works

### A. *SentiWordNet for Multilingual Sentiment Analysis*

Standard technology and existing sentiment analysis approach is a viable approach to sentiment analysis within a multilingual framework. We investigate methods to determine the polarity of sentences in a multilingual framework automatically. Sentiment Analysis within a multilingual context offers several challenges. Sentiment Analysis within a multilingual context provides several challenges. This first evaluation aimed to determine the performance of the different classification methods for English documents. The main benefit of the approach presented in this paper is the use of SentiWordNet as a lexical resource (Denecke, 2008).

### *B. Sentiment Analysis for Bangla Microblog*

Microblogging sites have become a viral source for publishing a massive amount of user-generated information. Sentiment analysis or opinion mining is the automatic extraction of opinions, emotions, and sentiments from texts. Sentiments, opinions, and emotions are subjective impressions and not facts, which are objective or neutral. We use SVM and MaxEnt and do a comparative analysis of these two machine learning algorithms' performance by experimenting with a combination of various sets of features. We assumed that tweets are subjective and hence did not deal with the neutral class. However, in the real world, objective tweets do not express any sentiment and fall into the neutral class (Chowdhury & Chowdhury, 2013).

### *C. Using Contextual Valence for Sentiment Detection*

The sentiment of a sentence or paragraph from Bangla text senses is different according to their parts of speech in the corresponding sentence. The senses may have nonzero scores for all the three categories, which indicates the corresponding word has each of the three opinion related properties to a certain degree. We propose a framework for sentiment detection based on valence analysis using SentiWordNet. Some Bangla texts mentioned from which detected sent. There is no problem with using WordNet and SentiWordNet for Bangla text (Hasan & Badiuzzaman, 2014).

### *D. Text Analysis with LSTM*

The traditional method is minimal, as it is unable to deal with vast amounts of data timely. Analysis text sentiment and RNN is a good model. (RNN) can take advantage of all previous words. Traditional RNN language model is going further in model generalization: instead of considering only the several previous words (parameter), the recursive weights represent short-term memory. LSTM through deliberate design to avoid long-term dependence, in practice, remember the long term information. For the RNN model, scientists put forward an idea named 'Attention,' which may lead to better results. LSTM also showed excellent performances in some application scenarios (Li & Qian, 2016).

### *E. Convolution Neural Network*

*S*entiment can is the opinion of people to-wards a specific interaction. Sentiment Analysis is the process of understanding people's opinions. Our proposed solution uses a Convolutional Neural Network which is a variant of an Artificial Neural Network. Their proposed system is not suitable for analyzing complex sentences. Pooling will result a feature vector for the individual feature map.

Moreover, we also compare our obtained accuracy with the accuracy of some other general Bangla sentiment classifiers. The number of comments in our dataset to train our model more effectively. We find our obtained accuracy is the highest in analyzing Bangla sentiment by our observation (Alam, Rahoman, & Azad, 2017).

### *F. Sentiment Analysis a Core Learning*

The researchers have taken several approaches to model human language in machine form pre-scented a sentiment detection technique using valency of a word. Machine learning algorithm needs the data represents a particular form; otherwise, the algorithms could not be implemented. It may be possible to reduce the length of the hash feature vector, and Accuracy is the measure of how correctly a classifier classifies data. Sentiments are part of the raw emotions of a text. They are developing a system that can read data from different sources like the web, hard drive, database, Etc. A way of extracting features from these data is described earlier (Hasan, Islam, & Hasan, 2019).

### *G. Recognition Techniques For Bangla Characters*

Bangla characters are involved in shape, very similar, and some of them can be written in multiple forms. Bangla script is also used in some other languages such as Meithei and Bishnupriya Manipuri, along with Bengali. The region sampling based method selects the regions which contain the most discriminating features of the image that describe the character. Converted the compound character to two or three primary characters to reduce the complexity of the feature set. It is also observed that CNN based methods gained more accuracy than other applied procedures. In addition to this, altering layers and filters can achieve more accuracy (Ghosh, Abedin, Chowdhury, & Yousuf, 2019).

### *H. Support Vector Machine(SVM) On Bangla Texts*

Bangla is an Indo-Aryan language primarily spoken by the Bengalis in South Asia. Two significant works of OCR contributing to research on Bangla primary characters involve a multistage approach MLP based classification technique. CMATERdb dataset has vast variations in sample grey-level images over each class with no fixed size. The significant advantage of SVM is that it is beneficial in high dimensional spaces, and the kernel method provides the decision function more adaptability. The datasets available for Bangla compound characters are tiny in size (Kibria, Ahmed, Firdawsi, & Yousuf, 2020).

## III. Methodology

### A. Data Set Description

This dataset text is preprocessed and tokenized into words. There is a total of 2210 readers and labeled by five different classes of sentiments. These classes are- 1. Very negative; 2. Negative, 3. Neutral, 4.Positive, 5. Very Positive.

Our LSTM model has been trained to show the percentages of the sentiment throughout the given sentence. Such as, every words in dataset has been categorized with five class sentiment mentioned above and these expressions are valued equal to some numerical value. Likewise, we trained the model to show Very negative = -2, Negative = -1, Neutral = 0, Positive = 1 and lastly, Very positive = 2.

### B. The Method Used to Explain

**Layer-wise Relevance Propagation (LRP)** is one of the most prominent methods in explainable machine learning (XML). The purpose of LRP is to provide an explanation of any neural network's output in the domain of its input (Lindwurm, 2019).

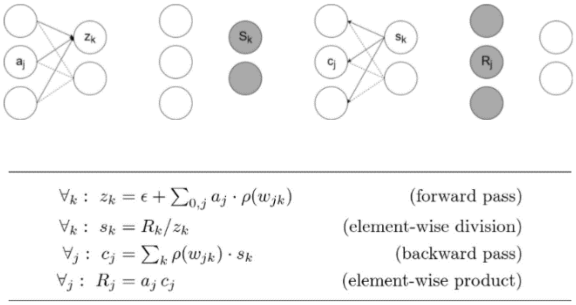
it uses the network weights and the neural activations created by the forward-pass to propagate the output back through the network up until the input layer. There, we can visualize which pixels really contributed to the output. We call the magnitude of the contribution of each pixel or intermediate neuron "relevance" values *R*.

LRP is a *conservative* technique, meaning the magnitude of any output *y* is conserved through the backpropagation process and is equal to the sum of the relevance map *R* of the input layer. This property holds for any consecutive layers' *j* and *k*, and by transitivity for the input and output layer.

From there on we go backwards through the network by following this basic LPR rule:

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$$

Here, *j* and *k* are two neurons of any consecutive layers. We already know the relevance *R* in the output layer, so we'll start from there and use this formula iteratively to calculate *R* for every neuron of the previous layer. *a* denotes the activation of the respective neuron, and *w* is the weight between the two neurons (P. Schwarzenau, 1998).



| | |
|---|---|
| $\forall_k: \; z_k = \epsilon + \sum_{0,j} a_j \cdot \rho(w_{jk})$ | (forward pass) |
| $\forall_k: \; s_k = R_k / z_k$ | (element-wise division) |
| $\forall_j: \; c_j = \sum_k \rho(w_{jk}) \cdot s_k$ | (backward pass) |
| $\forall_j: \; R_j = a_j \, c_j$ | (element-wise product) |

Table from "Layer-Wise Relevance Propagation: An Overview

This is the simplest LRP rule. Depending on your application, you will perhaps want to use different rules, which will be discussed later. All of them follow the same basic principle.

### C. The Method Used In Implementation

**Long short-term memory** (**LSTM**) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video).

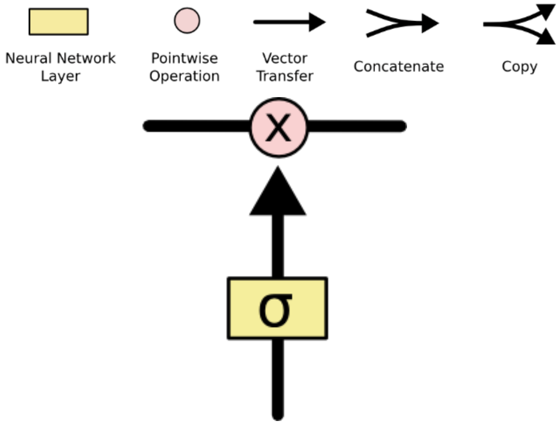


Figure 1: Necessary symbol to Understand LSTM

The sigmoid layer outputs numbers between 0-1 determine how much each component should be let through. Pink X gate is point-wise multiplication.

The core idea is this cell state Xt, it is changed slowly, with only minor linear interactions. It is very easy for information to flow along it unchanged. 2nd sigmoid decides how much information goes through. ”X” decides what info is to add to the cell state. Output gate controls what goes into output. Here, sigmoid is acting as a switch like 0 and 1. (Figure: 2)

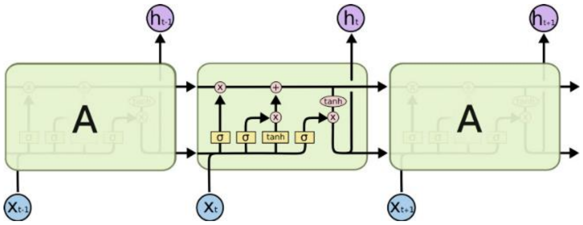


Figure 2: LSTM mechanism

A recurrent neural organization model we utilize a one covered up layer bi-directional LSTM (biLSTM), prepared on five-class sentiment forecast of expressions and sentences on the Stanford Sentiment Treebank film audits dataset (Socher, et al., 2013), as was utilized in past work on neural organization interpretability and made accessible by the creators. This model takes as information a grouping of words m1, m2,..., mT (just as this arrangement in switched request), where each word is spoken to by a word inserting of measurement 60 and has a shrouded layer size of 60. A careful model depiction can be found in the Appendix, and for subtleties on the preparation we allude to (Li , Chen, Hovy, & Jurafsky, 2016).

## IV. Result & Discussion

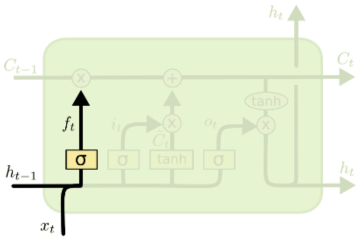
A pre-trained model was used from (Socher, et al., 2013) and as well as movie review treebank dataset was used to train the model. As a result of the importance disintegration for a picked target class, we initially get for each word inserting xt in an info sentence, a vector of pertinence esteems. To get a scalar word-level significance, we remind that we essentially summarize the correlations contained in that vector. Additionally, note that, per definition, the SA correlations are positive while LRP pertinent relationships are agreed upon [ras003]. In the following model, followed few steps those are- Decomposing Sentiment onto Words, Representative Words for a Sentiment, Validation of Word Relevance, Relevance Distribution over Sentence Length to make the model work.

Our model will give Bangla sentiment analysis by improving their model. For this first we give input to the system then it translates the Bangla into English by googletrans API (python) (“আমি ভাল আছি।” translated into “I’m fine.”). Not all translated words are acceptable by this model. That is why we went through few preprocessing steps to make it useable for the model. Firstly, we had to deconstructed words like “I’m” to “I ‘m”, then we made all upper cases words to lower case such as “I ‘m” to “i ‘m”. After making all lower case letter we have tokenized the words (“i ‘m fine.” To [“i”,”’m”,”fine”,”.”]). After than if we let our processed keywords to the model (Arras, Montavon, Muller, & Samek, 2019). then it shows it predicts class 4 which is a positive. By this we are getting 40 % precision on Bangla sentiment analysis.

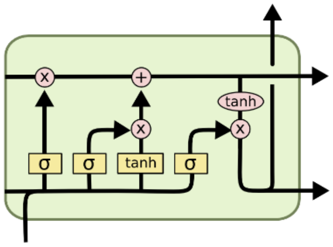
Overall, on five-class feeling forecast of full sentences (negative, negative, impartial, good, exceptionally good) the model accomplishes 46.3% precision, and for double grouping (good versus negative, overlooking nonpartisan sentences) the test precision is 82.9% (Arras, Montavon, Muller, & Samek, 2019).

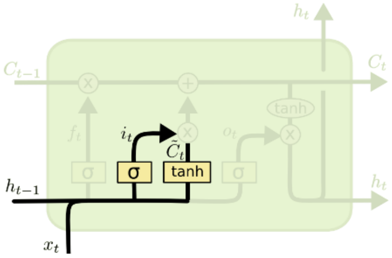
## V. Conclusion

In our paper, we have introduced a simple yet an effective strategy for extending the LRP procedure to recurrent architectures, such as LSTMs, by proposing a rule to back propagate the relevance through multiplicative interactions. We applied the extended LRP version to a bi-directional LSTM model for the sentiment prediction of sentences, demonstrating that the resulting word relevance’s trustworthy reveal words supporting the classifier’s decision for or against a specific class, and perform better than those obtained by a gradient-based decomposition. Our technique helps to understand and verifying the correct behavior of recurrent classifiers, and can detect important patterns in text datasets. Compared to other non-gradient based explanation methods, which rely e.g. on random sampling or on iterative representation occlusion, our technique is deterministic, and can be computed in one pass through the network. Moreover, our method is self-contained, in that it does not require to train an external classifier to deliver the explanations, these are obtained directly via the original classifier. Future work would include applying the proposed a technique to other recurrent architectures such as character-level models or GRUs, as well as to extractive summarization. Besides, our method is not restricted to the NLP domain, and might also be useful to other applications relying on recurrent architectures.



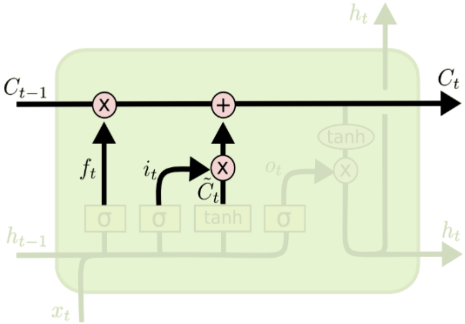
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \ + \ b_f\right)$$





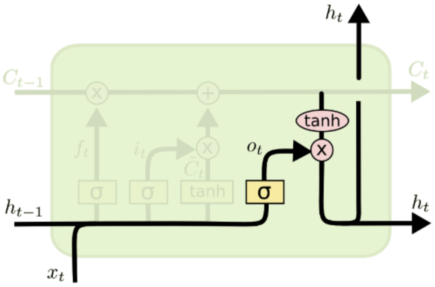
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$

$i_t$ decides what component is to be updated.
$C'_t$ provides change contents



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Updating the cell state



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

Decide what part of the cell state to output

Figure 3: LSTM explanations

## References

Alam, M., Rahoman, M.-M., & Azad, M. (2017). Sentiment Analysis for Bangla Sentences using Convolution Neural Network. Rangpur, Bangladesh: IEEE.

Arras, L., Montavon, G., Muller, K., & Samek, W. (2019). Explaining Recurrent Neural Network Predictions in Sentiment Analysis., (pp. 159-167). Germany.

Chowdhury, S., & Chowdhury, W. (2013). *Sentiment Analysis for Bangla Microblog Posts.* Dhaka: Brac University.

Denecke, K. (2008). Using SentiWordNet for Multilingual Sentimental Analysis.

Ghosh, Abedin, Chowdhury, & Yousuf. (2019). A Comphrehensive Review on Recognition Techniques for Bangla Handwritten Characters. Dhaka, Bangladesh: ICBSLP.

Hasan, K., & Badiuzzaman, M. (2014). *Sentiment Detection From Bangla Text using Contextual Valency Analysis.* Khulna, Bangladesh: ICCIT.

Hasan, M., Islam, I., & Hasan, K. (2019). Sentiment Analysis Using Out of Core Learning. *ICECCE.* Khulna, Bangladesh: IEEE.

Kibria, Ahmed, Firdawsi, & Yousuf. (2020). Bangla Compound Character Recognition using Support Vector Machine(SVM) on Advanced Feature Sets. Dhaka, Bangladesh: IEEE.

Kohlbrenner, M., Bauer, A., Nakajima, S., Binder, A., Samek, W., & Lapuschikn, S. (2020). Towards Best Practice in Explaining Neural. Germany.

Li , J., Chen, X., Hovy, E., & Jurafsky, D. (2016). Visualizing and Understanding Neural Models in NLP. *Conference of the North Amercia Chapter of Association for Computational Linguistics*, (pp. 681-691).

Li, D., & Qian, J. (2016). Taxt Sentiment Analysis Based on Long Short-Term Memory. *IEEE Conference.* Beijing, China: IEEE.

Lindwurm, E. (2019, December 15). *InDepth: Layer-Wise Relevance Propagation*. Retrieved from towards data science: https://towardsdatascience.com/indepth-layer-wise-relevance-propagation-340f95deb1ea

P. Schwarzenau, M. F. (1998, March). *A new method for the estimation of the onset of the lateralized readiness potential (LRP)*. Retrieved from Spring Link: https://link.springer.com/article/10.3758/BF03209421

Socher, Perelygin, Wu, J., Chuang, J., Manning, & Ng., C. (2013). Recursive Deep Models for Semantic Compositionality Over a sentiment Treebank. *Conference on Empricial Methods in Natural Language Processing(EMNLP)*, (pp. 1631-1642).