Gender Classification from Bangla Texts Using Deep Learning Approach

Adiba Haque(ID 17201113)
Anika Nahian Binte Kabir(ID 18101249)
Mayesha Monjur(ID 18101411)

1 Introduction

The automatic classification of gender based on cues is gaining popularity as the range of its application is expanding remarkably. Our project idea is to take innovative approaches in gender profiling from Bangla texts using deep learning models - Recurrent Convolutional Neural Network (RCNN), bi-LSTM and a more refined model by Bartle et al. Windowed Recurrent Convolutional Neural Network (WRCNN). Bengali language is the seventh most spoken language in the world. Working with this inherently complex and rich language will come with some challenges, however, our effort is to make available the user information of Bangla writers and bloggers for applications in social context, marketing, legal investigation and more. In the project we plan on using texts from Bangla microblogs such as Muktomona, Pachforon, Somewhereinbangla.net etc, texts and comments from Twitter, Google reviews and Facebook content purely in Bangla. We also plan to collect data from Bangla books, newspapers and magazines.

2 Background

Previously we have seen works on gender determination from text data in English and other languages. The aim to produce a similar project for Bangla text is the new take here. Inspired by the work in the paper: **Gender Classification with Deep Learning (Bartle et al.)**, our project applies Recurrent Convolutional Neural Network (RCNN), bi-LSTM and Windowed Recurrent Convolutional Neural Network (WRCNN), for gender classification.

3 Our idea and concept

In our experiment, we apply three different models on our dataset after textual pre-processing. In the first step we use a bi-LSTM and later, we shall compare its potential results with RCNN and windowed-RCNN. For RCNN, our first goal would be to create a left and right context vector and in order to achieve that, a bi-LSTM is used. By capturing the left and right contextual information in the form of vectors, we can reflect the true meaning of the word. The left context vector is constructed using the forward LSTM and the right context vector is created by the reverse LSTM of the bi-LSTM. Using the right and left contexts, a context vector for that particular word is created. A latent semantic vector is calculated by feeding each word context vector through the tanh activation function. Finally, all the latent semantic vectors are max-pooled together to give us a document vector. The document vector is fed through standard soft-max function to yield the output. Lastly, we use a modified version of RCNN, that is WRCNN, since a global approach does not necessarily yield desired results due to the presence of short -range dependencies. WRCNN takes smaller window vectors into consideration. Furthermore, it iterates through documents, sentences and words denoted by k,j and i respectively. Context vectors and latent semantic vectors are created by iterating through i^{th} word, j^{th} sentence and k^{th} document. Finally, the latent semantic vectors and maxpooled together, represented by the following equation:

$$y^{(4)}(d_k) = maxy^3(s_{j,k})$$

$$y^{(5)}(d_k) = softmax(W^5 y^{(4)}(d_k) + W^{(4)})$$

We employ precision, recall and F1-scores as an evaluation measure for the performance of our models.

4 Potential Challenges

An underlying problem with RCNN might be the way it treats the entire document as a training example. When blog posts or paragraphs from books consist of hundreds of words, the effectiveness of such Bidirectional RNN becomes less clear. Throughout the years, research progressed significantly for languages like English, Chinese, Arabic, etc. but not much development occurred for Bangla which is inherently morphologically complex. The presence of compound or joint words in Bangla vocabulary might be hard to tokenize and understand by the models. Furthermore, there can be significant vocabulary and grammatical differences between the data sources, also a mix of words from other languages since several words from other languages have entered and thus became a part of this language throughout the years. The blog datasets may also contain emoticons and additional characters in unrecoverable unicode format.Additionally, blog datasets being inherently biased may affect the overall performance and results.

5 Conclusion

Through our experiment, we have shown few models to classify author gender across several media using Bangla text. To obtain a baseline accuracy, bi-LSTM and an extension of the RCNN model was also proposed. An improvement was made by applying max pooling on sentences and this obtained a classification accuracy comparable to the previous works. Future work can be done by enhancing the neural networks as well as investigating other types of media.

6 References

1. Bartle, A., & Zheng J. (2015). Gender Classification with Deep Learning. Journal of Technical Report

2. Safara, F. (2020). An Author Gender Detection Method using Whale Optimization Algorithm and Artificial Neural Network. IEEE