Bangla Sentiment Analysis: An Unsupervised Hybrid Neural Network Approach

Quazi Adibur Rahman Adib,[1] Md. Sadman Sakib, [2] Kabbya Kantam Patwary,[3] Md. Humaion Kabir Mehedi,[4] Md Sabbir Hossain,[5] Annajiat Alim Rasel[6]

Department of Computer Science and Engineering, BRAC University

[1] quazi.adibur.rahman.adib@g.bracu.ac.bd, [2] md.sadman.sakib2@g.bracu.ac.bd, [3] kabbya.kantam.patwary@g.bracu.ac.bd, [4] h.k.mehedi11@gmail.com, [5] md.sabbir.hossain1@g.bracu.ac.bd, [6] annajiat@bracu.ac.bd

Abstract

The increased availability of smart devices and easy access to the internet has drawn many individuals to engage in different social media and express their opinions by reacting, commenting and sharing the contents. These interactions, therefore, produce a large pool of relatively ill-formed, casual and opinionated data which might be of great interest while working to analyze public sentiment. Being no exception, a reasonable amount of data has also been generated in Bangla language however it's pretty low compared to the data in English thus making it somewhat difficult to do sentiment analysis. Despite being marginally short on data, a few endeavors have been made for Bangla sentiment analysis mostly using techniques like CNN, LSTM, SVM and Naive Bayes on a supervised learning setup.(Irtiza Tripto and Eunus Ali 2018; Chowdhury and Chowdhury 2014; Sarkar and Bhowmick 2017) In order to bring a change to the way of addressing sentiment analysis, we're proposing an unsupervised hybrid neural network approach.

Initially, we have used One dimension Convolutional Neural Network (CNN) to extract features from the text. Subsequently, we used two unsupervised machine learning algorithms named K-Means Clustering and Gaussian Mixture algorithm to train our model. In this method, we have taken the advantage of a neural network as a feature extractor and used unsupervised machine learning algorithms as a classifier which marks this model as an unsupervised hybrid neural network. The dataset (Sazzed 2020) we used consisted of 8500 positive reviews and 3307 negative reviews. We divided it into a proportion of 75% and 25% for the purpose of training and testing respectively.

Among the two of the unsupervised clustering algorithms we used, the K-Means algorithm gained 86.92% accuracy and a score of 80.39% in the F1 metric. The Gaussian Mixture algorithm on the other hand obtained 92.24% accuracy and a score of 94.46% in the F1 metric.

References

Chowdhury, S.; and Chowdhury, W. 2014. Performing sentiment analysis in Bangla microblog posts. In *2014 International Conference on Informatics, Electronics Vision (ICIEV)*, 1–6.

Irtiza Tripto, N.; and Eunus Ali, M. 2018. Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 1–6.

Sarkar, K.; and Bhowmick, M. 2017. Sentiment polarity detection in bengali tweets using multinomial Naïve Bayes and support vector machines. In *2017 IEEE Calcutta Conference (CALCON)*, 31–36.

Sazzed, S. 2020. Cross-lingual sentiment classification in low-resource Bengali language. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, 50–60.