**Sentiment Analysis of Books and Its impact on children using Deep Learning**

Atif Ahnaf Chowdhury

Department of Computer Science

*International Islamic University Malaysia*

[*atifchy22@gmail.com*](mailto:atifchy22@gmail.com)

**Abstract – *Books are the most essential utility for human beings to enhance their knowledge. Using vulgar words in book is not new, from years people are using it but we were never concerned about the impact of it on a child. There are a lot of publisher outside who are involved printing books which contains vulgarism without knowing the side effects of those books on children. The reason for doing this research is to help the publisher and general people by identifying those words using the power of deep learning****.*

Keywords — Children Books, Vulgar words, Deep Learning, Sentiment Analysis, Naïve Bayes, LSTM, NLTK

# Introduction

Google says that book is a written or printed work consisting of pages glued or sewn together along one side and bound in covers. It can be full of vulgarism which gives a negative influence on the children. For example, Scottish author James Kalman saw his own book “How late it was how late” banned for its use of offensive language. Children are considered with the age range between seven to fifteen. The intention of this analysis is to find out the sentiment of the child’s book. Text does not only talk in format contents, but also attitudinal information, including emotional states.

Now a days Internet has been ruling all over the world and that made everyone so easy to get access of the file transfer protocol servers. Even by comparing 20 years from now, things have changed very dramatically, and this vast change can also give negative impacts on the child-readers mind. There is no existing research paper that calculates the vulgarism of a story book or poetry book. In this research paper, it is proven that a machine learning model can classify vulgarism from a text as an unsupervised data by giving the negative, positive & neutral sentiment and will give us the accuracy of the model. Naive Bayes & LSTM is being used in this research which can tell the sentiment of that text. There are many ways to do sentiment analysis.

# Previous Works

There are so many researches have done papers based on the textual data where a sentimental analysis has done for finding the vulgar words in the text-based file. Since the data are going to be used in this research is unlabeled data, therefore there are some research paper related to unlabeled data where unsupervised learning can be included. In this section, articles which have unsupervised learning and the machine learning techniques have been used are discussed. data by comparing descriptive sensory evaluation results about taste and two jjamppong ramen types. Authors have trained large number of text-based data using a skip-gram model for analyzing reviews for food evaluation to measure consumers’ acceptance or preferences. As a result, the experiment can be not only beneficial for evaluating the consumer’s reviews but also helpful for improving the marketplace. In this paper, authors have only stated two state of sense out of five sense. For the future work, sight and touch which are the most important sensory for food evaluation can be considered for analysis.

Caihong, Qiang, Hui, Shengqi, Nan and Zhengzheng (2018) used Artificial Neural Network (ANN) and Long-Short Term Memory (LSTM) models for forecasting hourly runoff discharges in Jingle hydrology station control catchment basin. Compared with flooding events simulation, ANN model is more sensitive that has many abnormal fluctuations, while LSTM model is more intelligence than ANN model. In this study, the runoff is changed in time-series that the data is time related. The ANN model is constructed by fitting the different characteristics of the current state and making prediction, while LSTM model not only take full advantage of the current data characteristics but also use its gate structure to decide to remember or forget the previous features. Instead of predicting the next hours runoff, in the future, its aimed to predict different length of time series with the entire sequence of data.

Cecilia Ovesdotter Alm, Dan Roth and Richard Sproat (2015) has provided an analysis on the textual data to emotion. The main goal of the research was to build a supervised machine learning model with SNoW learning architecture. Here authors wanted to classify the emotional affinity of sentences in the narrative domain of children’s fairy tales. There are two experiment was held on analyzing the textual data, for the first experiment, authors has built a classification model for a cross validation, and on the other experiment, they have split their emotional category into positive verse negative valence.

Kauter, Breesch and Hoste (2018) has worked on company specific sentiment analysis for financial news. To implement it a corpus of company-specific news articles which are marked by four annotators to create a standard. They used two coarse-grinded baseline systems: a lexicon-based approach outperforms both baselines and where the output is substantial to perfect agreement. They have also implemented Fine-grained sentiment annotation scheme to identify explicit and implicit expression of positive and negative sentiment. Their future work is the task of target attribution for topic-dependent sentiment analysis in financial texts.

Depti Mahajan, Radika patil and Varsha shankar has shown an improving approach upon the wordv2 skip gram model to build a better word representation using morphological rich languages. Here, authors has trained two models on the English corpus, one with 3 gram embeddings and another with 2-6 grams embeddings. Model is evaluated base on the similar words, word analogy and comparison of words. This is basically a modified version of skip-gram model.

Shrey Gupta and Armin Namavari (2017) has provided a analysis exploring word sense disambiguation (WSD) using two technique. There are several ways to compute senses. One better way is to use RNN. Using RNN model prediction can be easier because RNN will output a probability distribution over the next possible words. The results of our models are promising and should improve as more hand-labeled sense-tagged data becomes available.

For the future work, authors wanted to improve their application by using NLP technique such as machine translation and NER.

Alexander Pak and Patrick Paroubek (2010) has performed linguistic analysis of the collected corpus and explain discovered phenomena. They built a sentiment classifier that is able to determine emotions into three categories which are positive, negative and neutral. They have tested the impact of an n-gram order on the classifier’s performance. They also used Tree Tagger for POS-tagging and observed the difference in distributions among positive, negative and neutral set. Their data is also an unsupervised data here.

From the literature reviews above it is clearly seen that for sentiment analysis every author has classified the sentiments into positive or negative and whereas most of them used corpus as their refined data store. Some author got a good accuracy using Lexicons where as Kaihon (2018) got a very high accuracy using ANN and LSTM. In effect, for this project LSTM will be applied for the model building. However, this is challenging to detect vulgar words and its sentiment from a text-based file which is a book. Additionally, this project will has achieved better accuracy and is able to tell the impact of a book on children.

# Methodology

In this section, the process of getting data, cleaning the data, accuracy of LSTM & Naïve Bayes model and sentiment analysis is discussed in detail synchronously.

**3.1 Data Acquisition**

The data was collected from an online platform and the data which is chosen is an unsupervised data. The data itself came with a pdf format. The data is a story book for children.

**3.1.1 Changing format –** The data was taken in pdf format and then it was converted to text format.



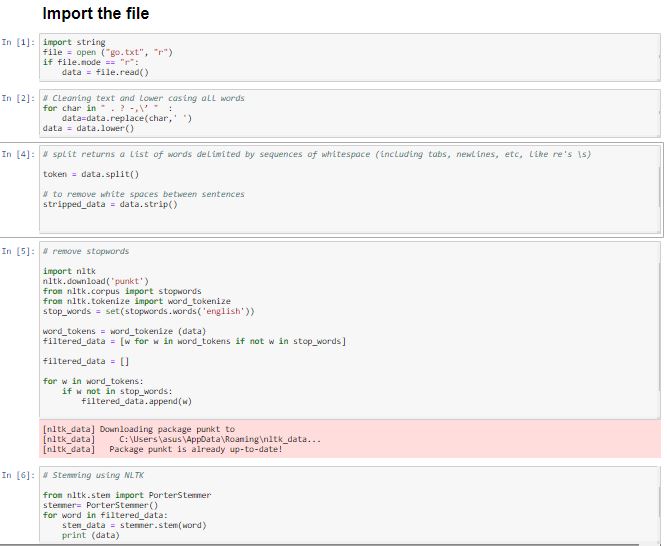


**Figure 1: Changing formats of text**

From these story books, we chose to work with “Go to the fuck to sleep”

**3.2 Data Preprocessing and Visualization**

In order to clean the text file pandas library was chosen to fit the text into a data frame. Firstly, all the texts were taken to lower case using python function and those texts were split using simple python function. And to clear the stop words and do the stemming, NLTK library was used. In figure 2, the steps are clearly shown. In figure 3, we have shown the specific words that are in the book. This book contains 227 distinct words. To get the visualization of the words that has been used in the book, wordCloud was imported and matplotlib library was used to see the plot of those words. After implementing these two libraries, the output was exactly as figure 4. It shows the most appeared words in the book which makes it easy to have a clear view of the book.



**Figure 2: Cleaning the text**



**Figure 3: Data frame of words**

****

**Figure 4: Visualization of words**

**3.3 Model**

**3.3.1 LSTM** Deep learning technique Long-Short Term Memory (LSTM) was considered for predicting the words. The TensorFlow library with keras was used for implementing LSTM model and the whole process was run using Notebook. The model was Train on 157 samples, validate on 40 samples, Total params was 96,337, Trainable params was 96,337 and Non-trainable params was 0. Total of three epochs gave the same accuracy which was 82.50%. After evaluating the model on the test set the loss was 67.2% and the accuracy was 62.9%.

**3.3.2 Naïve Bayes** is another deep learning technique which assumes that features of a measurement are independent of each other. We can simply take each feature separately and determine proportion of previous measurements that belong to class A that have the same value for this feature only. For our project, we took 300 subject documents and 300 object documents to train and test our model. The accuracy of the model was 80%.

After getting the accuracy Vader was used to test the sentiments. Vader is a parsimonious rule-based model for sentiment analysis of text. For the data that we chose, our model was able to tell that there is 18% negativity, 77.3% neutral and 4.7% positive.

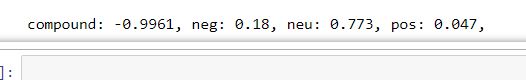
# Result and Analysis

In this section, the results yielded from the model will be showed and analyzed to find out the sentiments of text and impact on children.

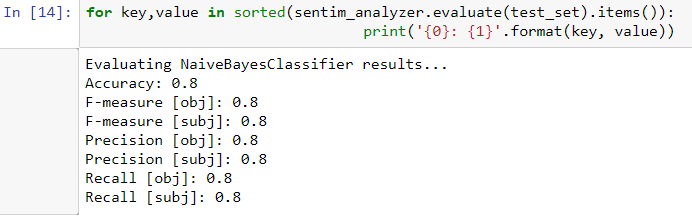
**Naïve Bayes & LSTM**

The Naïve Bayes model gave 80% accuracy and helped to define sentiment of text. Figure 5 shows the sentiments of our data that we chose for this research which states 18% negativity, 77.3% neutrality and 4.7% positivity. Figure 6 shows us the accuracy of Naïve Bayes model.

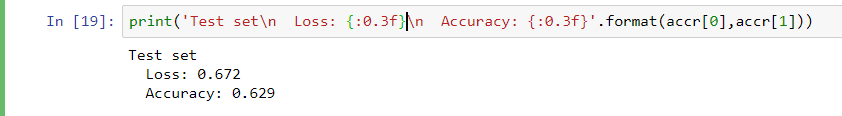
The LSTM model are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. In our project, our model gave us 62.9% accuracy which is quite high as figure 7 shows.



**Figure 5: Sentiments with Naïve Bayes**



**Figure 6: Accuracy of Naïve Bayes**



**Figure 7: Accuracy of LSTM model**

# Discussion

In this section, further discussion on the results were done for getting the accuracy using another machine learning algorithms and finding the sentiments.

At the very beginning, word2vec was implemented to get the sentiments of data. By using word2vec, classification of words can be easy to get. But we faced difficulties while getting the accuracy, the model gave us an error message saying that the weights of the words were initially sorted whereas we did not sort the weight of the data. Gensim library was used in this case. Bag of words (BOW) was also used to separate specific words and get the accuracy of the model. While using BOW , we saw some error saying that list strings cannot be sorted. Then it was decided to work with naïve bayes which did not give any error while building or training the model. It also gave the sentiments by using Vader NLTK which results are helping to categorize the books. Data was also fed to LSTM model which gave a accuracy of 62.9% as shown in figure 7.

# Conclusion

In this paper, we learnt more about Sentiment analysis of books and how it can make an impact on children and the society. It was found out that Naïve Bayes performed well to determine the sentiment analysis. Using the model, we can get a good accuracy on defining the bad or vulgar words and give a result, though it is difficult to accurately identify the exact sentiment all the time. The process for identifying sentiment analysis is based on which the model will learn by the data we put and then proceed with the sentiment analysis. These kinds of research can help understanding the quality of a book and its effects on our children. This can be transformed in an early warning system which can aid the society and specially for the parents to get aware about the children so that they can identify which books can be good for their children. This research can be really effective in a broad range if we can established in a common platform where everyone can easily know about our model and learn how to use this and implement it to determine the specific sentiment of the book that they are going to buy for their children. So, it will help their children in a way that they will not read something unexpected and they will be out of bad effects of any books, which will help a society and the country as well, when the young people will get good quality books to read on the time of their childhood, it will make a huge impact.

# Acknowledgment

The authors would like to acknowledge the contribution of Dr. Sharyar Wani for his constant support and suggestion throughout the research period.

# References

List of all the references is given *below*.

[1] Augustine Yongwhi Kim, JinGwanHa, Hoduk Choi, Hyeonjoon Moon. (2018). Automated Text Analysis Based on Skip-Gram Model for Food Evaluation in Predicting Consumer Acceptance.

.

[2] Caihong, Qiang, Hui, Shengqi, Nan and Zhengzheng . (2018). Deep Learning with a Long Short-Term Memory Networks Approach for Rainfall-Runoff Simulation.

[3] C, Dan Roth and Richard Sproat ecilia Ovesdotter Alm. (2005). Emotions from text: machine learning for text-based emotion prediction.

[4] Marjan Van de Kauter , Diane Breesch , Véronique Hoste. (2015). Fine-grained analysis of explicit and implicit sentiment in financial news articles.

[5] Depti mahajan , Radika patil, Varsha Shankar. (2018). Word2Vec using Character n-grams.

[6] Shrey Gupta, Armin Namavari. (2017). Word Sense Disambiguation Using Skip-Gram and LSTM Models.

[7 C´ıcero Nogueira dos Santos, Ma´ıra Gatt . (2014). Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts.

[8] . Svetlana Kiritchenk, Xiaodan Zhu, Saif M. Mohammad. (2014). Sentiment Analysis of Short Informal Texts.

[9] Bo Pang and Lillian Lee. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts.

[10] Alexander Pak, Patrick Paroubek. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining