TEXT SUMMARIZER

## Nagam Vinusha1, Panthangi Shiva Sowmya2,Naineni Anasri3

1Department of Computer Science and Engineering, SR University, Warangal, Telangana, India. 2Department of Computer Science and Engineering, SR University, Warangal, Telangana, India. 3Department of Computer Science and Engineering, SR University, Warangal, Telangana, India.

# Abstract

The most common problem in the present period is text summarization. Text summarization is a procedure that produces a concise, fluid, and, most importantly, accurate summary of a lengthy text material. With this strategy, we developed a model that will condense the text and produce a summary of our text data. Finding a condensed subset of the most crucial information from the complete collection and presenting it in a comprehensible style is the fundamental goal of automatic text summarization. It is quite helpful because more information that is desired may be read quickly. Whereas the human eye may overlook important information in your content, our software does not. To build this model, we have used seq2seq modeling for encoding and decoding the text and Long Short time memory(LSTM) for building the model. To train the model, we used the dataset which is collected from kaggle website. In our project we got an accuracy of 84.86% for the 1,00,000 records we have taken.

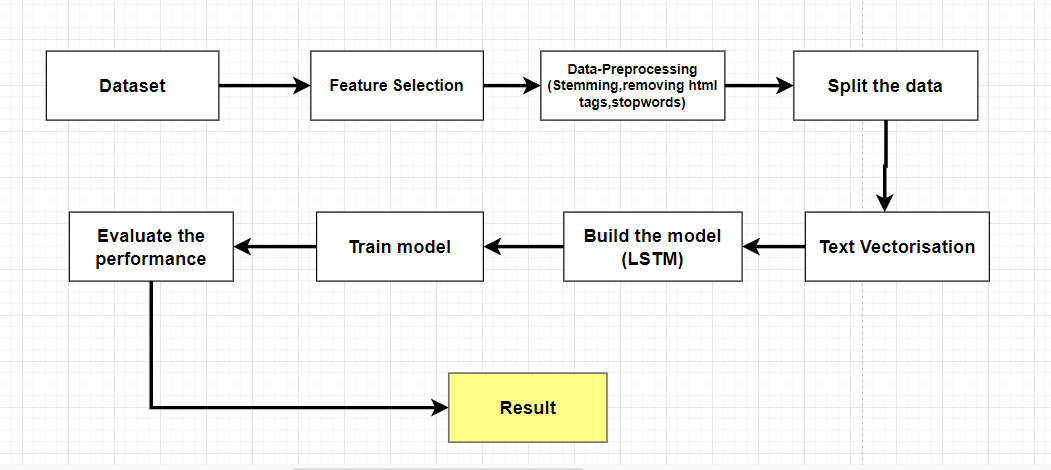
**Keywords:** Text Summarization, seq2seq model, Encoding and Decoding, LSTM

**1. INTRODUCTION**

Since the middle of the 20th century, text summarising research has been researched. Lun (1958) used word frequency diagrams as a statistical tool to explain the topic in public for the first time. There have been a lot of various strategies developed so far. There are single and multi-document summarizations depending on the document count. Meanwhile, the extractive and abstractive outcomes are based on the summary results.

The process of creating a concise, fluid, and, most importantly, accurate summary of a lengthy text content is known as text summarization. The fundamental goal of automatic text summarization is to be able to extract the most important information from a large body of text and display it in a way that is human readable. In several fields, automatic text summarization is applied. The field of news is one of the ones where Text Summarization is most frequently used. For instance, we read a lot of news items every day. The majority of the time when reading an article, we encounter many facts that may not be necessary to the story or that may not be that crucial for us to know. Text summary approaches in this situation enable us to reduce the size of the news piece by providing the essential information.

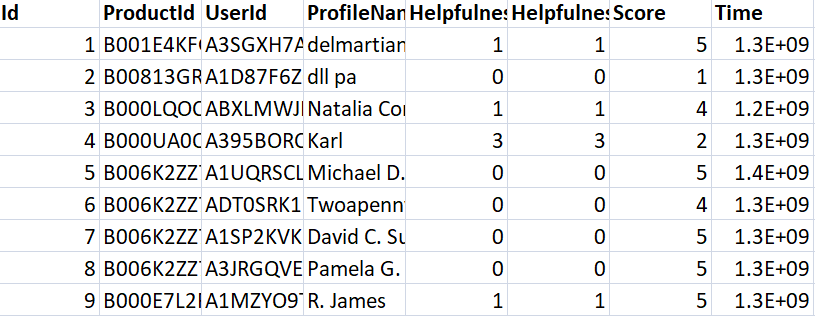
**2.METHODOLOGY**

****

**Figure1:Project Workflow**

**2.1 Dataset**

The dataset is downloaded from kaggle which was containing all the data regarding customer reviews on food .This data was extracted from the amazon. This dataset contains 5,68,454 customer reviews and 10 columns like text, summary, product id, userid, name etc. But we are taking 1,00,000 rows from dataset for training and testing the model. The columns text and summary is used in building the model.



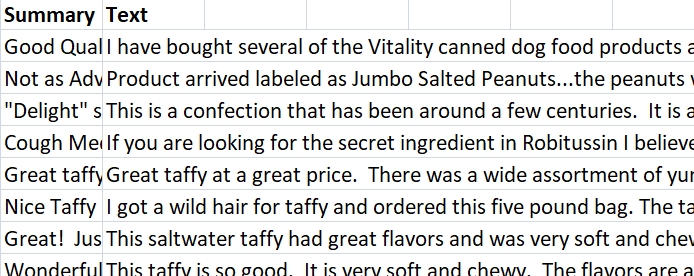


Figure 2: Sample data of Amazon fine food reviews dataset.

**2.2 Data Preprocessing**

Data pre-processing is a technique that is used to convert raw data into a clean dataset. The data is gathered from Kaggle which is in raw format (i.e., unbalanced data and unrequired columns present in it) which is not feasible for the computer to predict the summary of the given text. Pre- processing for this text data is determined below:

**2.3 Removing Unnecessary Columns:**

The columns like time, profile name, product id, user id ,score is not necessary for building the model. So, we are deleting all the unnecessaruy columns and using only text, summary columns. Dataset after deleting all the un-necessary columns.

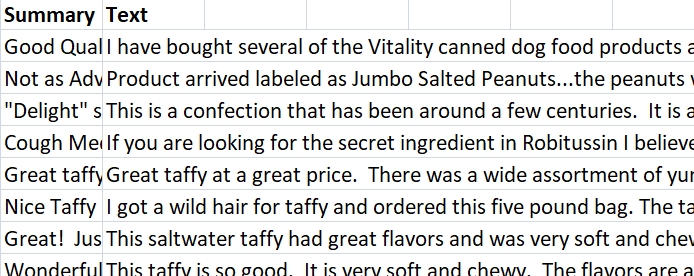


Figure3: Dataset after feature selection

**2.4 Removing Duplicate Data:**

The Values in the dataset may be duplicate so, there is a need to check for data that is repeating and remove it so that it does not affect the model training and the accuracy that it is giving.

**2.5 Removing Null Data:**

The Values in the dataset may be Null so, there is a need to check for data that is null or not, so that it does not affect the model training and the accuracy that it is giving.

**2.6 Removing Html tags:**

The reviews contains some html tags in between which is not useful and it also affect the model which decreases the accuracy. BeautifulSoup is used for removing html tags.

**2.7 Converting to Lower Case:**

Converting the text present in the reviews column into lower case is necessary since it will be helpful while converting the text data into vectors. These vectors will be later used for training the model and validating it using the validation data. All the text should be in same format for this process.

**2.8 Removal of Stop Words:**

Stop word removal is one of the most used preprocessing steps across different NLP applications. The idea is simply removing the words that occur commonly across all the documents in the corpus. Typically, articles and pronouns are generally classified as stop words. These words have no significance in some of the NLP tasks like information retrieval and classification, which means these words are not very discriminative. On the contrary, in some NLP applications stop word removal will have very little impact.

**2.9 Stemming**

Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. Stemming is important in natural language understanding (NLU) and natural language processing (NLP).

**2.10 Padding**

As we know all the neural networks needs to have the inputs that should be in similar shape and size. When we pre-process the texts and use the texts as an input for our Model. Where we know that we need to have the inputs with the same size, now here padding comes into picture. The inputs should be in same size at that time padding is necessary. The Padding method from Keras library is used to add zeroes to the formed vectors.

**2.11 Text Embedding**

This section talks about text encoding and decoding. To convert the data into integer format, we are using seq2seq model.

**Seq1Seq2 model :**

Seq2Seq model is a model that takes a stream of sentences as an input and outputs another stream of sentences. This can be seen in neural machine Translation where input sentences is one language and output sentences are translated versions of that language. Encoder and Decoder are the two main techniques used in seq2seq modeling.

**Encoder Model :**

Encoder model is used to encode or transform the input sentences and generate feedback after every step. This feedback can be an internal state i.e, hidden state or cell state if we are using the LSTM layer. Encoder models capture the vital information from the input sentences while maintaining the context throughout.In Neural Machine translation, out input language will be passed into the encoder model where it will capture the contexual information without modifying the meaning of the input sequence. Outputs from the encoder model are then passed into the decoder model to get the output sequences.

**Decoder Model :**

The decoder model is used to decode or predict the target sentences word by word. Decoder input data takes the input of target sentences and predicts the next word which is then fed into the next layer for the prediction.<start> to start the sentence and <End> to end the target sentence are the two word that help the model to know what will be the initial variable to predict the next word and the ending variable to know the ending of the sentence. While training the model, we first provide the word<start>, the model then predicts the next word that is the decoder data target. This word is then fed as input data for the next timestep to get the next word prediction.

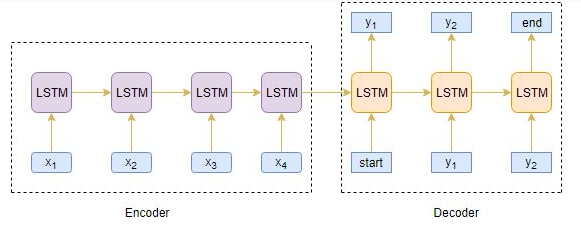


Figure 4: Architecture of Encoder-Decoder Model

**3. Model**

We are using Stacked LSTM containing 3 layers of LSTM stacked on top of each other. This will make our prediction much better. As per your requirement, you can have more also. Let’s understand our encoder model and decoder model.

**Encoder:** We will initialize the encoder input tensor using the ‘Input’ object. The expected shape of the batch will be 74 (maximum input length)-dimensions. Then we will create an ‘Embedding Layer’ which will have the total number of input words as the first argument and a shape of 500 which is the latent (hidden) dimension.

**LSTM:** Now we will create 3 stacked LSTM layers where the first LSTM layer will have input of encoder and like that create a continuous sequence of LSTM layers.

The LSTM layer will capture all the contextual information present in the input sequence. We will return hidden state output and also states i.e. hidden state and cell state after execution of every LSTM layer.

**Decoder:** Like Encoder we will initialize the decoder input tensor and then pass it to the only LSTM. Here, the decoder will also have the initial state where we will pass the hidden state and cell state values that we have obtained from the encoder’s LSTM layer.

**Attention Layer:** We will pass the encoder and decoder outputs into the attention layer and then we will concatenate attention layer outputs with the decoder outputs.

Now we will create our Dense Layer that is the output layer for our model. It will have the shape of the total number of target words and a softmax activation function.

**4. Results**

* **No. Of Epochs=10**

From split data 80% of training data, we made a machine model to predict. The machine model predicts the other 20 % of data as data testing to see how our model work. We use 3 stacked Long Short-Term Memory to make the machine model good enough to predict.

* **Accuracy:**

The results of this machine model are having 84.86% accuracy. It is explained from 20% data, that is classified for testing in which the training data is used to predict the machine correct around the trained model.

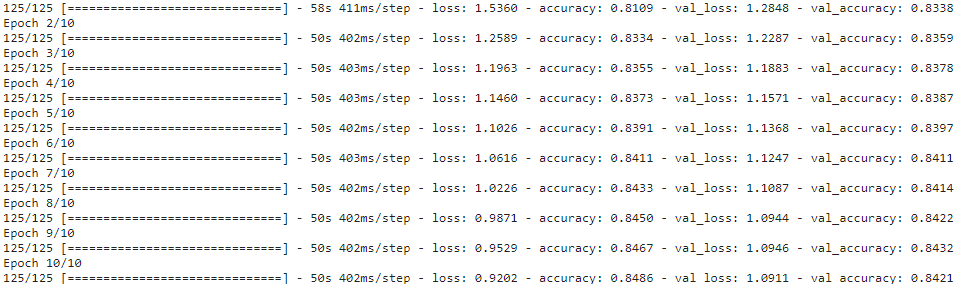


Figure 4: accuracy scores

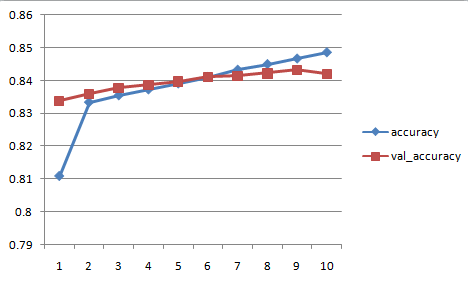
****

Figure5: accuracy Score Graph

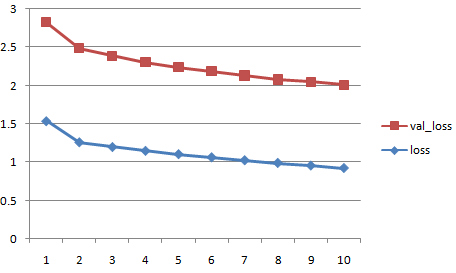
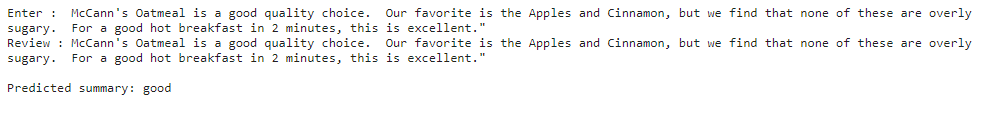
****

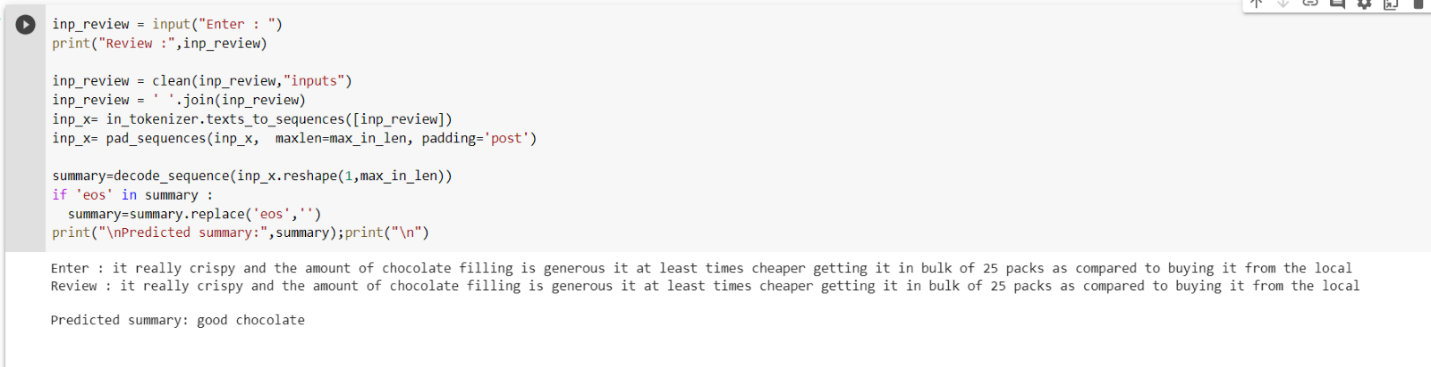
Figure6: Loss Value Graph

**Testing the model**

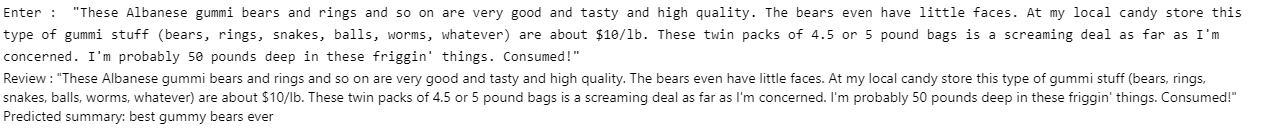
**Output 1:**

****

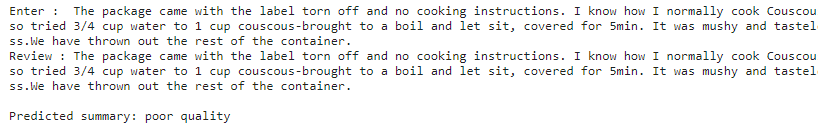
**Output 2:**

****

**Output 3:**

****

**Output 4:**

**+**

**Figure 9: Output**

**5.Conclusion**

In the new digital era where large amounts of datasets are budding on the globe, to analyze the same and extract the result is getting popular. While building the paper, we studied a lot of research in a statistical manner to understand the need and greed of visualizing the dataset and processing techniques applied on the text data. For the same, much research has been taken into count. By using seq2seqmodel with Lstm, we can predict the summarized text /review. So that the user can able to know the review without wasting time.The proposed model able to predict the summarized text for the given input with 84.86% accuracy.

**5.REFERENCES:**

[1] Dutta, Sushama Rani, et al. "IoT in Autism Detection in Its Early Stages." Internet of

Things: Enabling Technologies, Security and Social Implications. Springer, Singapore, 2021. 47-58.

[2] Dutta, Sushama Rani, Sujoy Datta, and Monideepa Roy. "Using cogency and machine learning for autism detection from a preliminary symptom." 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE,

2019.

[3] Dutta, Sushama Rani, et al. "A Machine Learning-Based Method for Autism Diagnosis

Assistance in Children." 2017 International Conference on Information Technology (ICIT). IEEE, 2017.

[4] Giri, Soumyajit, Sushama Rani, et al. "A real time epidemic alert generation system for

rural areas using WBANs and kiosks." 2017 International Conference on Information Technology (ICIT). IEEE, 2017.

[5] Dutta, Sushama Rani, and Monideepa Roy. "EDOT: Context-aware tracking of similar data patterns of patients for faster diagnoses." 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT). IEEE, 2017.

[6] Dutta, Sushama Rani, and Monideepa Roy. "Providing context-aware healthcare services

using circular geofencing technique." 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, 2016.]

[7] Dutta, Sushama Rani, et al.” Handling Scarcity of Data in Autism Identification using Binary Imputation Method” International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1, May 2019

[8] Donalek, Ciro. "Supervised and unsupervised learning." Astronomy Colloquia. USA. 2011.

[9] Kukkar, Ashima, and Rajni Mohana. "An optimization technique for unsupervised automatic extractive bug report summarization." International Conference on Innovative Computing and Communications. Springer, Singapore, 2019.

[10] Huai, Beibei, et al. "Mining Intentions to Improve Bug Report Summarization." SEKE.

Vol. 2018. 2018.

[11] Angel Hern ́andez-Casta ̃neda et al. “Extractive Automatic Text Summarization Based on Lexical- Semantic Keywords”. In: IEEE Access 8 (2020), pp. 49896–49907.