Abstractive Summarization of Product Reviews with Sentiment Analysis

Bhairavi Sameer Shah[1], Michelle Elizabeth[2], Diya Liza Varghese[3], Theres Mary Jose[4], Aysha Fymin Majeed[5]

U.G Student, Department of Computer Science, Govt. Model Engineering College, Kochi, Kerala, India[1,2,3,4]

Assistant Professor, Department of Computer Science,Govt. Model Engineering College, Kochi, Kerala, India[5]

ABSTRACT: E-commerce websites allow customers to leave reviews for various products. There are usually hundreds of reviews for a single product, each review could be lengthy and repetitive. A customer would find it difficult to make a well-informed decision after reading all the reviews. Therefore, automatic review summarization has a huge potential to help customers by providing an authentic summary of the reviews found online on the E-commerce sites. We propose the method of abstractive summarization, which provides more accurate summaries and are closer to human generated summaries. The system also provides the general sentiment of the summaries generated which will help the customers make a decision quickly. The sentiment of the summaries would help the customers know the tone of the text.

KEYWORDS: Abstractive summarization; Sentiment analysis; Natural language processing; Neural Networks; Sequence-to-sequence with attention mechanism

I. Introduction

Information is always available in plenty in the modern era of the Internet. Anything that happens in any part of the world reaches the other side in the blink of an eye, and hence, access to information is the last thing that you will need to worry about. According to Google Search Engine results stats, around 1 billion articles, posts or information of any kind is released for a single focused keyword every day. And, for a consumer, information available in the web will remain incomprehensible, unless it is transformed in a way that would help them understand the data in the most comprehensive way possible.

The best way to provide the right amount of information that a user requires is to limit the content into precise and accurate points. Providing the gist of the enormous content will reduce the overhead of processing unwanted information for consumers, and the difficulty of providing meaningful data for the providers. A summary can be used, which is a subtle way of representing a lot of information in a minimal form. This seems to be profitable on a lot of platforms. Now, making a summary manually would require the user to read all the content again, which nullifies the sole aim of summaries. This is where an automatic summary generator comes into the picture.

Automatic summary generators provide you with a summary of the input text that you provide to the system. The automated summary generator can find its application in different forms, in educational fields, content creation, E-commerce, marketing, etc. As long as huge amounts of information needs to be processed in daily life, an automatic summarizer will also be pertinent.

II. Related work

In [2], a multi-view abstractive summarization model that jointly considered semantics and sentiment was examined. The proposed multi-view model, contains a basic summary component which performs sentence compression using encoder-decoder and a sentiment component which uses Sentiment Embedding to extract and represent the sentiment features automatically, and Sentiment Memory to capture the sentiment dynamics during the process. In [3], the Deep Neural Network model, namely Sequence to Sequence model, is applied to summarise research articles by considering the introduction and conclusion sections. Temporal attention mechanism is used to subdue the occurrence of repetitive words or sentences in summaries. According to [4], there are two approaches for performing abstractive summarization of multiple documents. First approach is phrase selection and merging, which uses a linear optimization method to obtain a summary. The second method is the Semantic Information Extraction Approach which uses BSU basic semantic network which depicts semantic information. The projected system in [5] identifies the number of positive and negative opinions of each aspect in online reviews. The system preprocesses data and performs sentence and aspect extraction. Stop word removal, stemming and POS tagging are the processes in data preprocessing. Naive Bayes algorithm is used to identify opinions. [6] introduces the decoder attention mechanism in the pointer-generator network to improve the accuracy and fluency of the summary. The multi-hop attention mechanism is introduced to improve the copy probability distribution. The coverage vector, used as a new input to compute the attention mechanism, tells the model which words it has focused on before so that fragment duplication is avoided. The proposed dual encoding model in [7] consists of a primary encoder, a secondary encoder, and a decoder. It conducts the primary encoder and decoder as the standard attentional encoder–decoder model. The secondary encoder is based on the input and the previously produced output, and generates a new context vector as an additional input of the decoder. Semantic role labelling is used in [8] to obtain the predicate argument structure. It is extracted to represent text semantically and semantically similar predicate argument structures are grouped using hybrid approach of K-means and agglomerative hierarchical clustering. In [9], two sets of summaries of the same data were obtained, one by automatic text summarization and the other by manually producing summaries. Automatic summaries were obtained by using fuzzy method and vector approach. The summaries produced by fuzzy method were much more acceptable and understandable compared to the ones produced by vector approach. [10] uses a clustered genetic semantic graph approach for multi-document abstractive summarization. The semantic graph is constructed by ensuring the graph vertices represent the predicate argument structures (PASs), extracted automatically by employing semantic role labelling (SRL) and the edges of graph correspond to semantic similarity weight determined from PAS-to-PAS semantic similarity, and PAS-to-document relationship.

III. Proposed methodology and discussion

Automatic summarization has great potential in the E-commerce field, to help customers make a well-informed decision while making purchases online. Automatic review summarization can provide a consolidated summary of the reviews found on the E-commerce sites. The project uses a Sequence to Sequence (Seq2Seq) model with Attention Mechanism which consists of an Encoder, Decoder, and Attention layer to perform abstractive summarization of user reviews. A naive bayes model is used to perform sentiment analysis.

A. *Data collection and preprocessing*

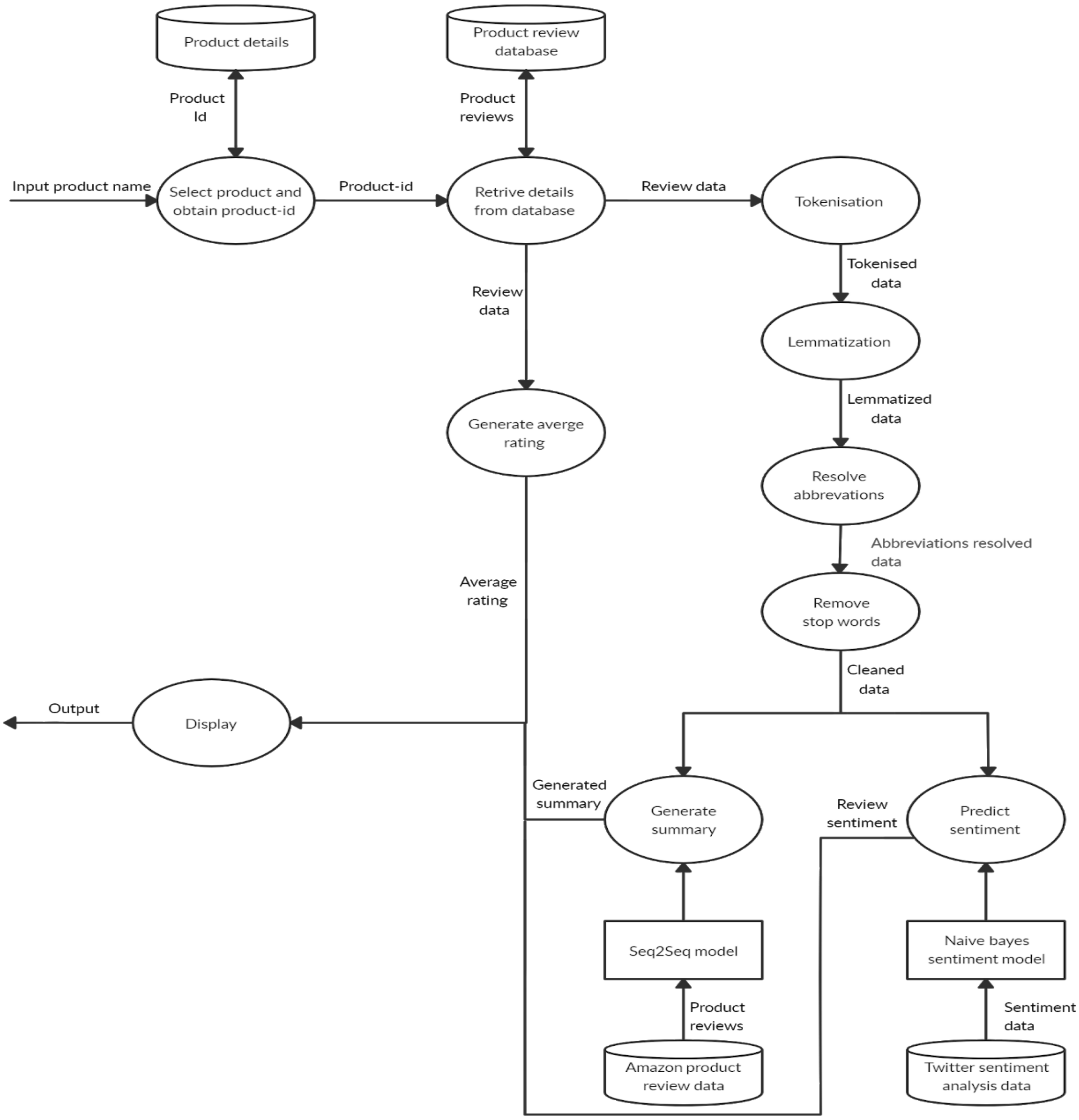
Before training both the models, each of the input datasets are preprocessed. Sms language and slang words are mapped using a dictionary of abbreviations to make the text more understandable. Short forms are replaced with full forms using a contraction mapping. The input text is then cleaned to remove noise by following the preprocessing steps which includes POS tagging, lemmatization, removal of stop words, removal of punctuations. In the case of summarization, a fixed length of review text as well as summary text is chosen based on the frequency of rare words in each of them, respectively. The rest of the review-summary sets outside of this condition are discarded. Rows with null values are removed from the dataset. The review text and summary text tokenizers are then fitted on this cleaned text using a Tokenizer, which are also used as input to the Seq2Seq model. In the case of sentiment analysis also, tokenization is performed on the tweets to obtain a vectorized form which is given as input to the naive bayes model.

B. *Training the machine learning models*

1) Seq2Seq model with Attention mechanism to generate summary.

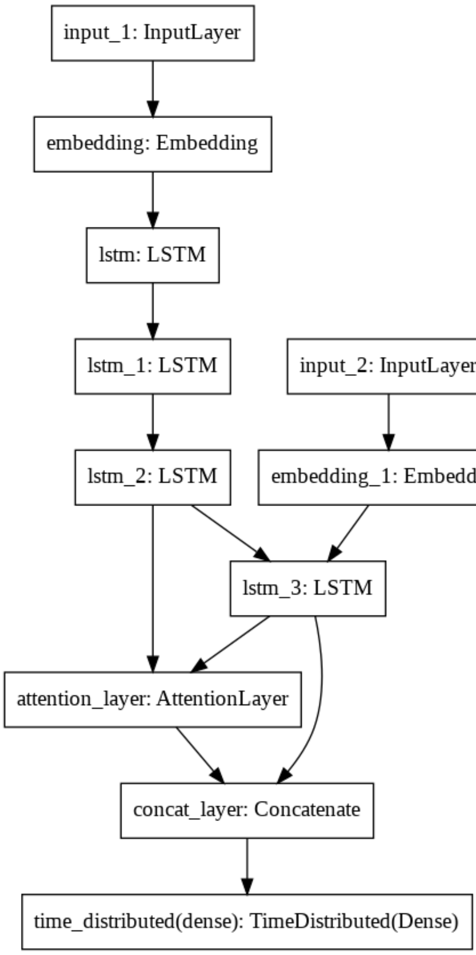
The Seq2Seq model is trained with preprocessed reviews and summaries as described earlier. The model has an encoder and a decoder which will help in producing an output that is of different length than the input. An encoder Long Short Term Memory model (LSTM) takes the input word by word in each time cycle and captures the contextual information in the input. The hidden state and cell state of the last time step is used to initialise the decoder. The decoder reads the entire target sequence word-by-word and predicts the same sequence offset by one timestep. The attention layer aids in predicting a word by looking at a few specific parts of the sequence only, rather than the entire sequence.

2) Naive Bayes model to find the sentiment of the reviews.
The sentiment analysis dataset, which contains tweets and their associated sentiment (positive{1} or negative{0}) is preprocessed and is used to train a Naive Bayes model. The model is tested with preprocessed testing data to obtain the sentiment of the input text and calculate its accuracy.



C. Summary Generation & Sentiment Analysis

The reviews available for a product are obtained from the dataset. This data is then preprocessed and given as input to the Seq2Seq model to generate the summaries. Extractive approach is used to fetch the top 5 summaries from those generated by the model to be displayed to the user. The preprocessed reviews are also sent to the Naive Bayes model to find the sentiment of each of the reviews. The percentage of positive and negative sentiments predicted are displayed to the user, thus providing the general sentiment obtained from all of the reviews.



IV. EXPERIMENTAL RESULT

A data set containing nearly 6,00,000 reviews of Amazon products and their summaries are used to train the sequence to sequence model. The NLTK corpus of twitter sentiment analysis data that contains tweets and their sentiment is used with various models such as Linear Regression, Naive Bayes, Random Forest, SVM and K-Nearest Neighbors from which the Naive Bayes model is chosen, which showed the best result among them. The sentiment model proved to have an accuracy of about 99% while testing.

	Predicted Positives	Predicted Negatives
Actual Positive	TP	FN
Actual Negative	FP	TN

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

The reviews are also passed into the trained deep learning model for summary generation, which is the Sequence to Sequence (Seq2Seq) model with Attention Mechanism to perform abstractive summarization of user reviews. A total of 10,905, 992 trainable parameters are used in the model. The model is trained for a total of 5 epochs with each epoch taking a training time of 990s on average. The total training time was around 1 hour 53mins.

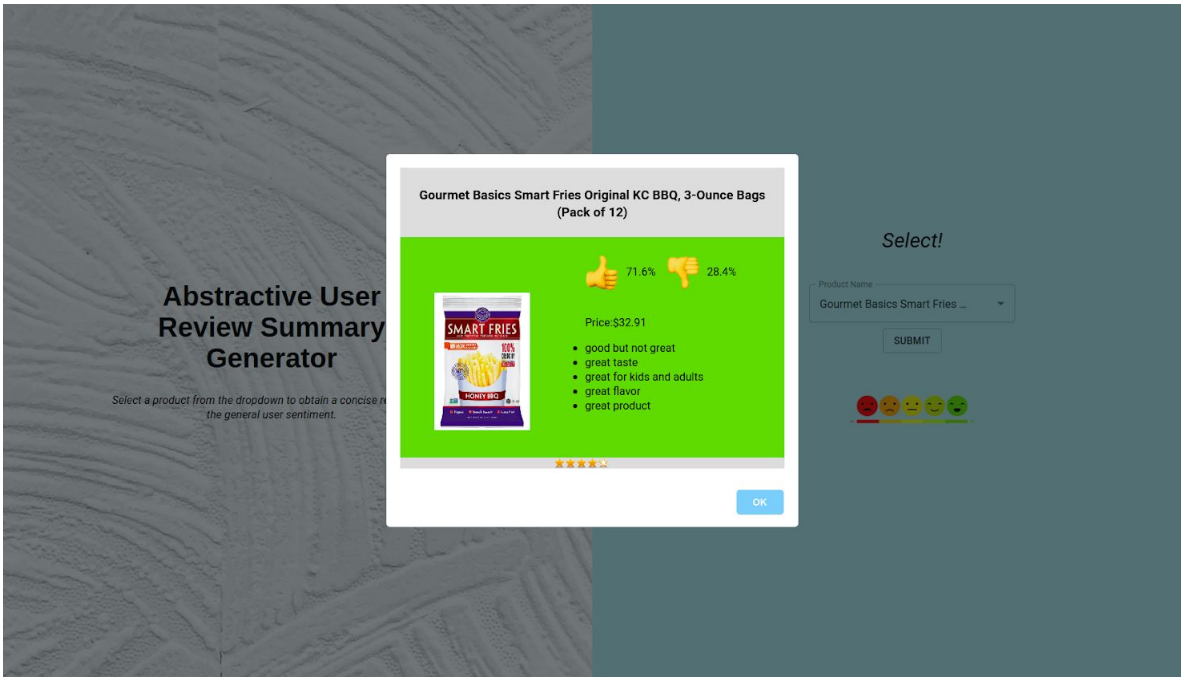
Preprocessed review text	Original summary	Predicted summary
fantastic coffee fraction price freeze dried keurig cup made company tasteless coffee bean matter name coffee cannot imagine paying dollar cup donut house cup made folgers joke	great coffee	great coffee
coconut water years back recall tasting like bad could even swallow one pack try add pineapple juice one containers remaining see makes difference wow	absolutely disgusting	tastes like plastic
got taste test bubble fizz taste chalk like bad sell taste sad	not very good	not good
found brand fact dark chocolate made far best keurig continue order particular brand	best ever	best chocolate ever
like jelly bellys think appreciate belly flops surprise flavors get worth every penny opinion size shape affect flavor still taste great	best flops ever	awesome
honestly know really get overly concerned ph balance electrolytes keep handy husband long mile bike trips us like tastes would rather drink gatorade also know silky smooth taste tastes like water wanted	do not know about health issues but tastes fine	do not know about health issues
cannot say enough good things snackwell fudge drizzle caramel popcorn remember snackwell products first introduced good real cookies come long way packs decent serving popcorn plenty crunchy caramel chocolate overwhelming great balance already purchased	highly addictive	good snack

Example preprocessed review text, original summary, and predicted summary

Even though the summaries generated are not exactly the same as the original, they still are comparable and in some cases even prove to be better.

e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 5, May 2020 ||



The summaries, positive and negative sentiment percentage, and the average rating is displayed along with product name, image and price.

V. Conclusion and Future Work

A user review consolidation system has been developed using the method of abstractive summarization to obtain a concise form of the lengthy product reviews available on e-commerce websites along with additional features such as overall rating and sentiment analysis. The Naive Bayes classifier for sentiment analysis provided an accuracy of more than 99%. The abstractive summarizer has been modelled as a Sequence-to-Sequence model which uses attention mechanism to improve accuracy. The model gives fairly good summaries of the reviews. The model is trained using reviews for a selected dataset.

The generalization capability of the deep learning model can be enhanced with an increase in the training dataset size. Web scraping can be introduced to perform summarization on real-time data. Implementing Bi-Directional LSTM which is capable of capturing the context from both the directions could result in a better context vector. The beam search strategy can be applied for decoding the test sequence instead of using the greedy approach (argmax). Pointer-generator networks and coverage mechanisms can be implemented in the model to further improve the summary generation capability of the model.

References

1. Chetana Badgujar, Vimla Jethani and Tushar Ghorpade, ‘Abstractive Summarization using Graph Based Methods’, Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018), IEEE 2018.
2. Moye Chen, Lei Li, and Wei Liu, ‘A Multi-View Abstractive Summarization Model Jointly Considering Semantics and Sentiment’, Proceedings of CCIS, 2018.
3. Tooba Siddiqui and Jawwad Ahmed Shamsi, ‘Generating Abstractive Summaries Using Sequence to Sequence Attention Model’, International Conference on Frontiers of Information Technology (FIT), 2018.
4. Ranjitha N S and Dr. Jagadish S Kallimani, ‘Abstractive Multi-Document Summarization’.
5. A. Jeyapriya and C.S.Kanimozhi Selvi, ‘Extracting Aspects and Mining Opinions in Product Reviews using Supervised Learning Algorithm’, IEEE Sponsored 2nd International Conference On Electronics And Communication Systems (ICECS 2015).
6. Wenbo Nie, Wei Zhang, Xinle Li and Yao Yu, ‘An Abstractive Summarizer Based on Improved Pointer-Generator Network’.

e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 5, May 2020 ||

7. Kaichun Yao, Libo Zhang , Dawei Du , Tiejian Luo, Lili Tao, and Yanjun Wu, 'Dual Encoding for Abstractive Text Summarization', IEEE Transactions on Cybernetics.
8. Alshaina S, Ansamma John and Aneesh G Nath, 'Multi-document Abstractive Summarization Based on Predicate Argument Structure', 978-1-5386-3864-4/17/$31.00, IEEE 2017.
9. Farshad Kiyoumarsi, 'Evaluation Of Automatic Text Summarization Based On Human Summaries', 2nd Global Conference on Linguistics and Foreign Language Teaching LINELT-2014.
10. Atif Khan, Naomie Salim and Haleem Farman, 'Clustered Genetic Semantic Graph Approach for Multi-document Abstractive Summarization', 978-1-4673-8753-8/16/$31.00, IEEE 2016.
11. Shuai Wang, Xiang Zhao, Bo Li, Bin Ge and Daquan Tang, 'Integrating Extractive and Abstractive Models for Long Text Summarization', 6th International Congress on Big Data, IEEE 2017.
12. Siddhartha Banerjee, Prasenjit Mitra and Kazunari Sugiyama, 'Multi Document Abstractive summarization using ILP Based Multi Sentence Compression'.
13. Alexander M. Rush, Sumit Chopra and Jason Weston 'A Neural Attention Model for Abstractive Sentence Summarization', Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI-2015).
14. Jianwei Niu, Huan Chen, Qingjuan Zhao, Limin Sun and Mohammed Atiquzzaman, 'Multi-Document Abstractive Summarization using Chunk-graph and Recurrent Neural Network', IEEE ICC 2017 SAC Symposium Big Data Networking Track.
15. Hao Xu, Yanan Cao, Ruipeng Jia, Yanbing Liu, Jianlong Tan "Sequence Generative Adversarial Network for Long Text Summarization", 30th International Conference on Tools with Artificial Intelligence, IEEE 2018.
16. Harsha Dave and Shree Jaswal, 'Multiple Text Document Summarization System using Hybrid Summarization Technique', 1st International Conference on Next Generation Computing Technologies, 2015.
17. Amazon review dataset, http://jmcauley.ucsd.edu/data/amazon/
18. Twitter sentiment analysis dataset, NLTK, http://www.nltk.org/howto/corpus.html