A SUMMARIZATION APPROACH FOR EVENT CENTERED BENGALI TWEETS

Under the Supervision of
Dr. Dwijen Rudrapal, Asst. Prof., CSE Dept.
Presented by :-
CHANCHAL KHATUA

August 17, 2023

Introduction

- Text summarization is a way to summarize a single document or multi-document texts and return a concise amount of text, a paragraph most likely from where we can have minimum idea about the whole content of the document.
- One of the most popular social media platform is Twitter. It contains mainly text data. The huge amounts of text data are generated by twitter's users in one day in various languages.

Problem Statement

- Mainly the Event based tweet summarization based on English tweets.
- Due to lack of Bengali Twitter Datasets, Bengali tweets are collected based on particular event or hashtag and created avg. 16 tokens of gist summary in each event.

Related Work

Table: 1 Related Work

Ref.	Proposed model	Event Categories	Result
[1]	SVM, Naive Bayes	Opinion Analysis on Sport event	86% accuracy in SVM
[2]	Hidden Markov Models	Sports Event	25% higher than base line approach
[3]	OntoDSumm	10 Disaster Events	66 % ROUGE-1 F1 score
[4]	EndSUM	6 Disaster Events	56% ROUGE-1 F1 score
[5]	ClusterRank, COWTS, FreqSum, LexRank, LSA,LUHN, MEAD, Sumbasic	5 Emergency Events with 1000 tweets	53% ROUGE-1 F1 score in LUHN
[6]	BERT	COVID-19 and UK election tweets	20% ROUGE-1 F1 score

Datasets Collection Procedure

- Bengali tweets are collected from Twitter API based on trending events or keywords.
- Remove unnecessary symbol, URLs, hashtags, extra space, single quote, double quote and punctuation from the text.
- Save all tweets of each event as one input line of Encoder input.
- Summary of tweets each event is created as gold summary.

Dataset Description

- In the datasets, total 46 events are present, and every event contain in average 600 tokens and summary contain average 16 token tweets.
- Dataset contain 57,637 total word with, 9924 unique words.

Dataset Description

Table 2: Details Description of Dataset.

Parameter	**Details**
No of events	46
Total tweets	5094
Min tweets of an event	6
Max tweets in an event	604
Total words	57637
Unique words	9924
Maximum input length of event	6227
Minimum input length of event	63

Word Count Frequency of Original & summary

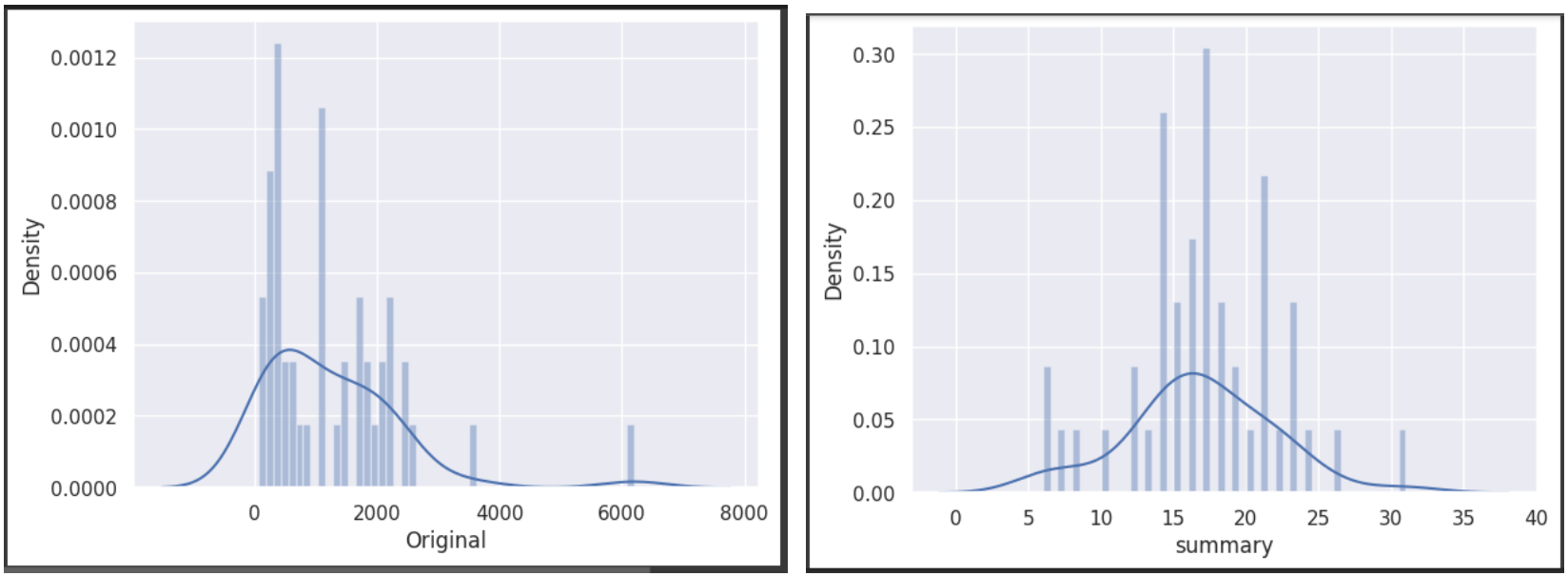

Fig 1: Word Count Frequency of Event in Original Text & Summary

Statistics of Dataset

Table 3: Categories of Event with Distribution.

Categories of Event	Number of Event	Distribution of Event	Number of Tweet	Distribution of Tweet
Tweets of famous person	8	15.39%	510	9.26%
Protest	6	13.04%	1363	24.76%
Sports	9	19.57%	935	16.98%
Disaster	5	10.87%	576	10.46 %
Entertainment	3	6.52%	535	6.41%
Politics	7	15.22%	872	15.84%
Other	8	15.39%	485	8.81%

Annotation Challenges

- One difficulty is dealing with tweets in different languages. It is a time-consuming process to filter out tweets in languages other than the desired language.
- Another difficulty is dealing with various symbols that are used in tweets. Symbols can vary greatly, and it can be challenging to identify and remove them.
- Creating gold summaries manually can also be challenging. It can be difficult to determine which tweets are the most important or common for a particular event.

Proposed Architecture

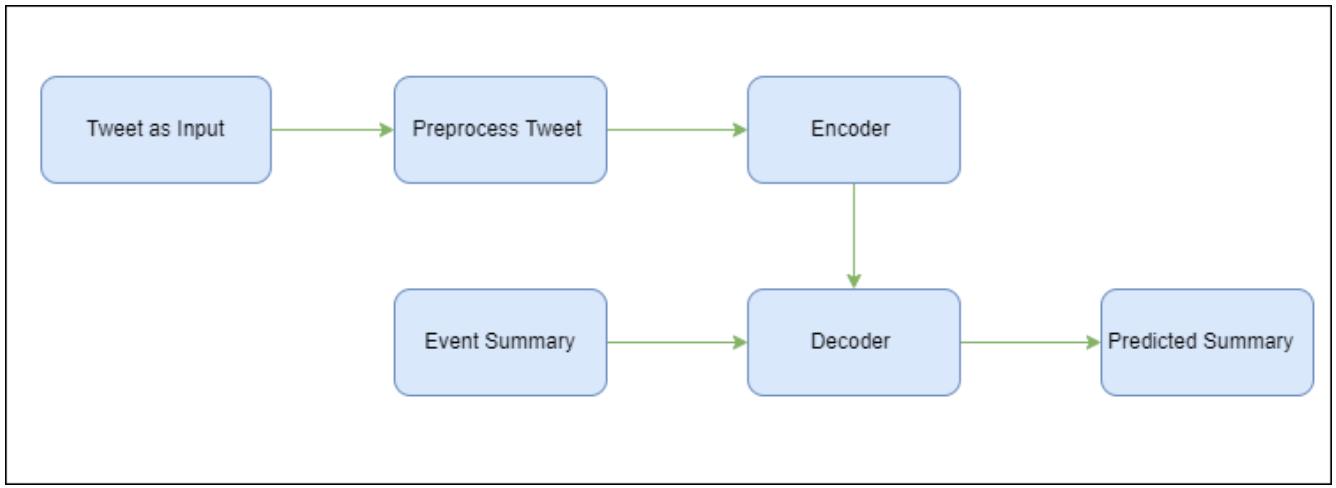

Fig 2: Proposed Architecture of Experiment

Proposed Model

- The model is built to take integer sequences as input and provide summary in the same format. Encoder and decoder with LSTM are the two fundamental parts of the model.
- A context vector is produced by the encoder after it has processed the input sequence. The decoder can use this context vector to create the output sequence by summarizing the input sequence.
- A dense vector representation of each word in the input sequence is created by the embedding layer of the encoder.
- An LSTM layer appears after the embedding layer, processing the embedded input sequence and producing a series of hidden states and a final hidden state.

Proposed Model

- The decoder uses the context vector produced by the encoder to produce the output sequence.
- Along with this, the decoder also has an embedding layer that turns each word in the target sequence into a dense vector representation.
- The final hidden state of the encoder LSTM layer is used to initialize the decoder LSTM layer.
- The decoder LSTM evaluates the embedded input sequence and generates a series of hidden states that are passed to a dense layer to predict the subsequent word in the target sequence.

Result

Performance of this experiment is compared with two exiting datasets in Table 4. The row A1 (proposed) of the table shows performance of the proposed model, and approach(A2)[9] is a Bengali abstractive news summarization approach based on encoder decoder model with LSTM on "BANS" dataset. In other approach(A3), [8] 36 news article tweets collected for the summarization and its ROUGE-2 score is better than the proposed model. Overall, performance of proposed is good among these approaches.

Performance Comparison

Table 4: Performance Comparison

Approach	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
A1	**0.71**	0.39	0.27	**0.72**
A2	0.30	Not Given	**0.31**	0.30
A3	0.66	**0.54**	Not Given	Not Given

Discussion

Figure 3 show that when original summary length is small, then generated summary is more accurate. But when original summary length is long, then model produce inconsistent summary. Most Probably the training dataset doesn't have enough examples of longer summaries for a particular category of events, then the model may not have learned to capture the nuances and complexities of longer summaries for that category.

Summary Comparison

Actual Summary: ব্রাভোকে জরিমানা ও সন্দেহ আলামিনের বোলিং অ্যাকশনে দ্বিতীয় টেস্টে নেই গেইল টসে জিতে ফিল্ডিংয়ে বাংলাদেশ ওয়েস্ট ইন্ডিজ 380 রানে অলআউট

Predicated Summary: জরিমানা সন্দেহ আলামিনের বোলিং অ্যাকশনে টেস্টে টেস্টে গেইল টসে ফিল্ডিংয়ে বাংলাদেশ বাংলাদেশ রানে অলআউট ওয়েস্ট ওয়েস্ট ইন

Actual Summary: বাংলাদেশ প্রধানমন্ত্রী শেখ হাসিনা প্রথম পাতাল মেট্রোরেল নির্মাণ কাজ উদ্বোধন করলেন হিরো আলমকে হারিয়ে দেওয়া হয়েছে আওয়ামী লীগ সরকারের পদত্যাগ দাবি সমাবেশে একই দিনে পাশাপাশি স্থানে আওয়ামীলীগ বিএনপির সমাবেশ

Predicated Summary: বাংলাদেশ প্রধানমন্ত্রী শেখ হাসিনা পাতাল পাতাল মেট্রোরেল নির্মাণ কাজ উদ্বোধন করলেন করলেন হিরো আলমকে হারিয়ে দেওয়া হয়েছে। আওয়ামী লীগ সরকারের পাশাপাশি স্থানে আওয়ামীলীগ সমাবেশ

Actual Summary: পাকিস্তানের নয় এশিয়া কাপ

Predicated Summary: পাকিস্তানের নয় এশিয়া কাপ

Actual Summary: দেশে শৈত্যপ্রবাহ নেই জানালো আবহাওয়া দপ্তর

Predicated Summary: দেশে শৈত্যপ্রবাহ নেই জানালো আবহাওয়া দপ্তর

Fig 3: Summary Comparison

Conclusion & Future Work

- The model's performance on this dataset is confident. Although some inconstancy is happening to generate big length summary.
- This experiment also have some limitation like can not classify event automatically, can't process real time tweet stream.

References

1 N Vijay Kumar and M Janga Reddy. Factual instance tweet summarization and opinion analysis of sport competition. In Soft Computing and Signal Processing: Proceedings of ICSCSP 2018, Volume 2, pages 153–162. Springer, 2019

2 Deepayan Chakrabarti and Kunal Punera. Event summarization using tweets. In Proceedings of the International AAAI Conference on Web and Social Media, volume 5, pages 66–73, 2011.

3 Piyush Kumar Garg, Roshni Chakraborty, and Sourav Kumar Dandapat. Ontorealsumm:Ontology based real-time tweet summarization. arXiv preprint arXiv:2201.06545, 2022.

4 Piyush Kumar Garg, Roshni Chakraborty, and Sourav Kumar Dandapat. Endsum: Entropy and diversity based disaster tweet summarization. arXiv preprint arXiv:2203.01188, 2022.

5 Gupta, A., Chugh, D., & Katarya, R. (2022). Automated news summarization using transformers. In Sustainable Advanced Computing (pp. 249-259). Springer, Singapore.

6 oumi Dutta, Vibhash Chandra, Kanav Mehra, Sujata Ghatak, Asit Kumar Das, and Saptarshi Ghosh. Summarizing microblogs during emergency events: A comparison of extractive summarization algorithms. In Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 2, pages 859–872. Springer, 2019.

7 man Munire Bilal, Bo Wang, Adam Tsakalidis, Dong Nguyen, Rob Procter, and Maria Liakata. Template-based abstractive microblog opinion summarization. Transactions of the Association for Computational Linguistics, 10:1229–1248, 2022.

8 oshni Chakraborty, Maitry Bhavsar, Sourav Kumar Dandapat, and Joydeep Chandra. Tweet summarization of news articles: An objective ordering-based perspective. IEEE Transactions on Computational Social Systems, 6(4):761–777, 2019.

9 Prithwiraj Bhattacharjee, Avi Mallick, and Md Saiful Islam. Bengali abstractive news summarization (bans): a neural attention approach. In Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020, pages 41–51. Springer, 2021.

THANK YOU !