A report on

Extraction-based Text Summarization of News Text using BERT Embedding

Introduction

In recent decades, the advancement in communication media and informatics has caused the huge increase in text data base that needs to be summarized and analysed. In this digital era, data is generated every second and the amount of data one can process is growing exponentially. This leads to the invention towards development of automatic text summarization for the frequent evaluation of the huge text within a very short span of time. Text Summarization is a technique used in NLP to shorten long pieces of text into summaries from resources like books, news articles, web pages, conference papers etc. Therefore, summarization is a short representation of the article which contains the key matters of the entire dataset. There are two methods of text summarization: Extractive and Abstractive [1, 3, 5]. Extractive summarization generates ranks and scores of sentences in the given text based on certain metrics and then extracts important information by collecting the few of top-ranking sentences. Abstractive summarization works based on rephrasing of the text to produce short representation rather than picking up the top important sentences and words from the given text itself.

This project, we present the study on extractive summarization for news. Extractive text summarization using the conventional method was a challenging task due to poor performance of the sentence embedding models. However, with the invention of modern advanced deep learning based embedding models like transformer-based BERT, the text processing in NLP has become more practical. In this project, the use BERT sentence embedding with various supervised learning models has been demonstrated for the extractive text summarization for news articles taken from news dataset. For extraction summarization with supervised learning models, first, binary class of labelled dataset is created to represent the summarization of the news article. The implementation has been experimented with Cornell Newsroom dataset Created by Grusky et al., 2018.

Methodology

BERT Embedding:

Embeddings are one of the most important aspects in the natural language processing. It is vector representation of list of words also known as sentence vectors. Google research team introduces a neural network architecture called transformer [2] which outperformed the conventional models for the sentence embedding. The detailed architecture of Transformer model is given below:

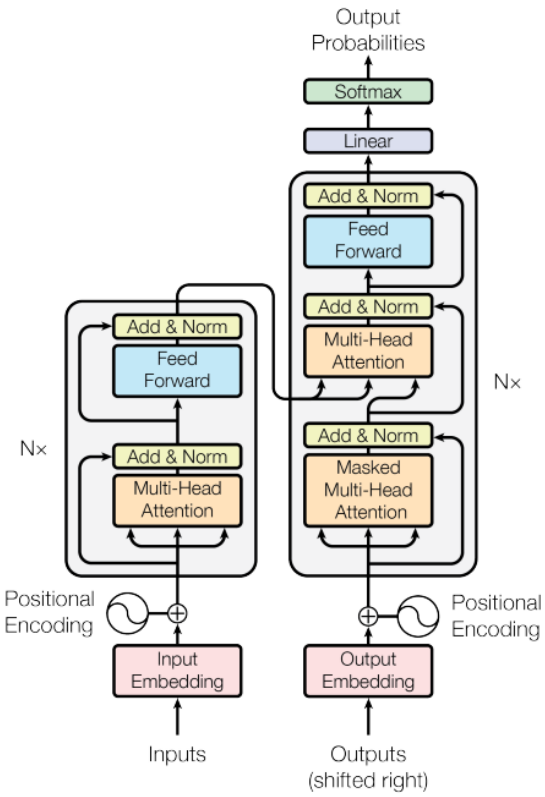


Figure 1: Transformer model architecture [2]

BERT works based on the mechanism of Transformer, an attention mechanism. It captures the contextual relations between words in the given article. It has two components: encoder and decoder. Encoder encodes the text inputs into low order vectors. On the other side, the decoder decodes back to the text which is the prediction for the given input. For the generation of the language model using the BERT, only encoder part of BERT is required in out text summarization process.

Sentence tokenization using Spacy:

All the processing in NLP requires the input text in token forms, therefore, sentence tokenization is the first step in the text processing pipeline. In this project, Spacy library has been used for cleaning the text and converting to tokens. The language processing pipeline using Spacy always depends on the statistical model and its capabilities.

Dataset Description

The dataset used in this Project is the Cornell Newsroom dataset Created by Grusky et al. in 2018. This dataset contains 1.3 million articles and summaries written by authors and editors in the newsrooms of 38 major publications. The summaries are obtained from search and social metadata between 1998 and 2017 in English language. First, the given JSON files for the train data and the test data are loaded and converted to PANDAS data frame. The dataset contains both extractive and the abstractive summaries. News articles with only extractive summaries are kept and rest are filtered out. After filtering, total number of samples in the train dataset are

332131 whereas the test dataset contains 36165 samples. The sentence embedder based on BERT can not be applied with given number of samples due to limitations with the computational resources. Therefore, the train dataset is truncated to 5000 samples and the test data is truncated to 500 samples.

Data Processing:

The data processing includes the filtrations to keep only "Extractive" followed by tokenization using Spacy. Then the cleaned text data is passed through the BERT embedder to obtain the embeddings. Then, a document label is added to each sentence by finding the cosine similarity between the summary sentences and article sentences. The sentences with high similarity are labelled as 1 and the other are labelled as 0.

Exploratory Data Analysis:

Here, the prepared dataset is analysed to understand the length, size, distribution, and quality of the dataset. Followings are the exploratory data analysis results:

1) Shape of the processed dataset:

```
Shape of xTrain (151173, 1538)
Shape of Train_doc_label (151173,)
Shape of yTrain (151173,)
Shape of xTest (14702, 1538)
Shape of Test_doc_label (14702,)
Shape of yTest (14702,)
```

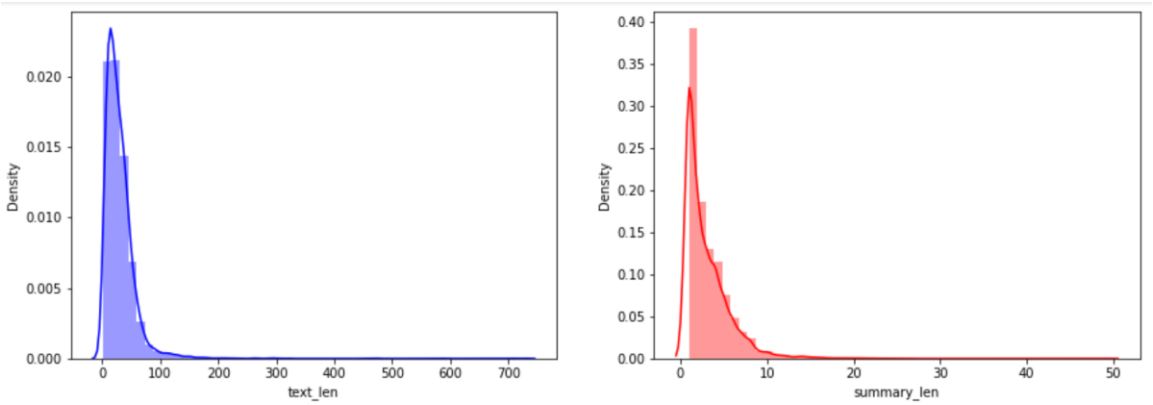
2) Mean length of the sentences in the text and the summaries:

	mean
Article	30.235
Summary	3.038

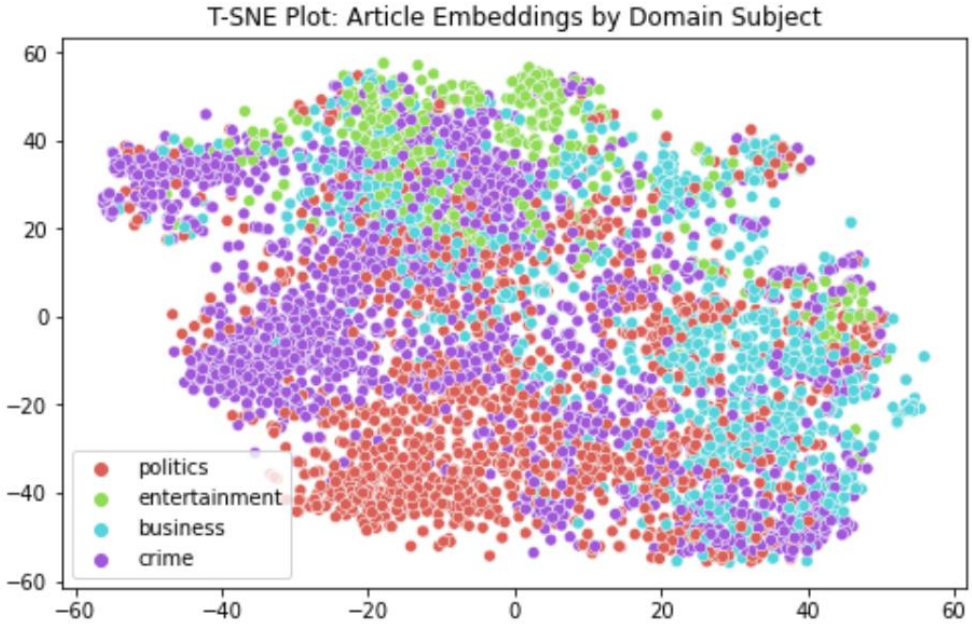
3) Statistical description of the dataset

	compression	coverage	density
count	5000.000000	5000.000000	5000.000000
mean	15.808952	0.966611	30.413089
std	23.386748	0.041583	21.966912
min	1.000000	0.547170	8.190476
25%	5.417964	0.953782	15.149351
50%	9.515877	0.978723	24.500000
75%	18.607794	1.000000	37.921840
max	687.200000	1.000000	281.062977

4) Distributions of the sentence lengths:

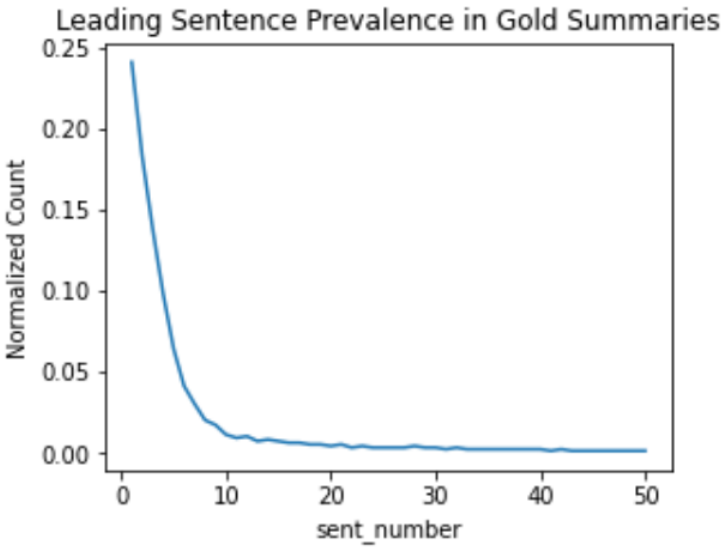


5) Embedding Visualization by Domain Subjects:



6) Description of the leading sentences:

sent_number	Normalized Count	Cumulative
1	0.241	0.241
2	0.183	0.424
3	0.138	0.562
4	0.098	0.660
5	0.064	0.724
6	0.041	0.765
7	0.030	0.795
8	0.020	0.815
9	0.017	0.832
10	0.011	0.843



Evaluation metrics:

In this project, Recall-Oriented Understudy for Gisting Evalution (ROUGE) [3] has been used. It is a set of different metrics like ROUGE-N, ROUGE-L, ROUGE- S etc. Here, we have used only ROUGE-1 and ROUGE- L with Precision, Recall and F1-score for the evaluation and comparison of the implemented models with the baseline models.

Implementation

Machine learning model and results:

1) **TextRank:** The TextRank model is implemented as the baseline model for the comparison results with all other. Textrank is a graph based ranking model which is used to find most important sentences in the text. Classification performance summaries are presented below:

S. N.	ModelID	recall	precision	F1-Score
1	TR:Rouge1	0.27691057	0.315885442	0.254634915
2	TR:RougeL	0.20086087	0.232472694	0.18556245

2) **Logistic Regression (LR):** Logistic regression works very well for binary classification problems. Here, Logistics regression model is implemented to predict the similarity index which may be either 0 or 1. Two case studies are presented under different conditions: default model and the Elastic Net model with hyperparameter tuning.
   a) **Default model (LR_default):** Logistic model with default parameters is trained and the classification results are obtained. The classification report in form of confusion matrix is given below:

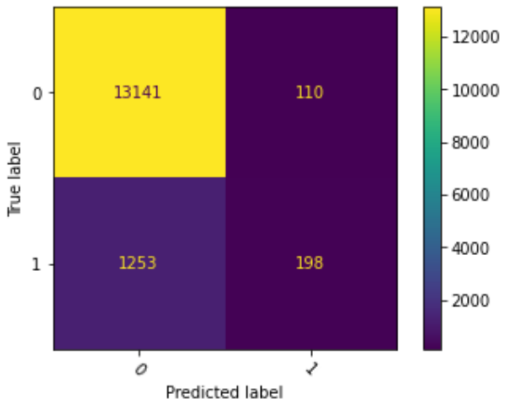


Figure 2: Classification report by default model of LR.

b) **Elastic net model (LR_elasticNet)**: It uses both L1 and L2 regularizations. Hence, model was tuned with the hyperparameter of L1 ratio as [0.25, 0.5, 0.75, 1]. Another hyperparameter that determines the strength of the regularization is called C, which is the inverse of regularization strength. The space for the C-values was chosen as [0.25, 0.5, 1, 2, 4]. The higher the values of C correspond to less regularization. The ROUGE-1 and ROUGE-L matrices were used for evaluation. Classification report for the LR_elasticNet is given below

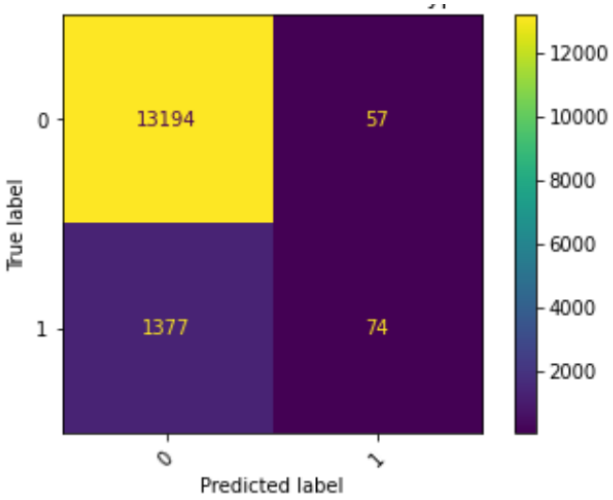


Figure 3: Classification report (confusion matrix) for LR_elsticNet.

Summaries of results for the two case studies:

S. N.	ModelID	recall	precision	F1_Score
1	LR_default:Rouge1	0.494898475	0.704380201	0.513244205
2	LR_default:RougeL	0.468242615	0.666323126	0.48664521
3	LR_elasticNet:Rouge1	0.508568711	0.727450582	0.532298757
4	LR_elasticNet:RougeL	0.48561015	0.693253405	0.508973611

3) **Neural Networks (NN):** Neural network is the one of the most popular methods for machine learning based modelling and data analysis [4]. It is implemented with Relu and sigmoid activation functions. Adam optimizer is used for parameter optimisation with cross entropy loss. Four case studies are presented with different model architectures. For each case, ROUGE-1 and ROUGE-L performances are evaluated.

   a) **Single hidden layer (NN_16):** NN model with single hidden layer having 16 nodes was implemented and trained for 50 epochs. The model details and the training loss curve are given below:

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_3 (Dense)             (None, 16)                24624

 dense_4 (Dense)             (None, 1)                 17

=================================================================
Total params: 24,641
Trainable params: 24,641
Non-trainable params: 0
_________________________________________________________________
None
```

Figure 4: Model Summary for NN_16

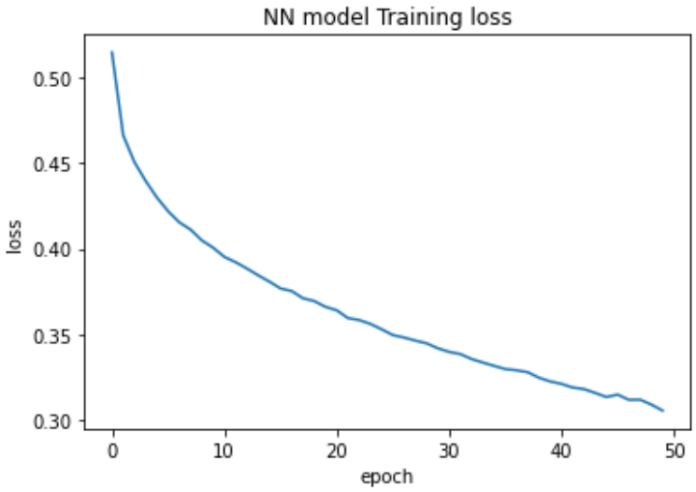


Figure 5: NN model training loss curve

   b) **Two hidden layers:** NN models with two hidden layers have different combinations of neural nodes have been implemented as follows:

```
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 dense (Dense)                (None, 32)                49248

 dense_1 (Dense)              (None, 8)                 264

 dense_2 (Dense)              (None, 1)                 9

=================================================================
Total params: 49,521
Trainable params: 49,521
Non-trainable params: 0
```

Figure 6: Model summary for NN_32_8

```
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 dense_8 (Dense)              (None, 32)                49248

 dense_9 (Dense)              (None, 16)                528

 dense_10 (Dense)             (None, 1)                 17

=================================================================
Total params: 49,793
Trainable params: 49,793
Non-trainable params: 0
```

Figure 7: Model summary for NN_32_16.

```
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 dense_5 (Dense)              (None, 64)                98496

 dense_6 (Dense)              (None, 16)                1040

 dense_7 (Dense)              (None, 1)                 17

=================================================================
Total params: 99,553
Trainable params: 99,553
Non-trainable params: 0
```

Figure 8: Model summary for NN_64_16.

Results summary for all cases of NN models is presented in the following table

S. N.	ModelID	recall	precision	F1-Score
1	NN_16:Rouge1	0.555342057	0.746306077	0.565704822
2	NN_16:RougeL	0.537553059	0.718951647	0.547235296
3	NN_32_8:Rouge1	0.560763474	0.729521174	0.563155645
4	NN_32_8:RougeL	0.541521775	0.70060453	0.543299784
5	NN_32_16:Rouge1	0.557103403	0.744142115	0.565558498
6	NN_32_16:RougeL	0.53827103	0.715306563	0.545916449
7	NN_64_16:Rouge1	0.556960206	0.736682986	0.565115908
8	NN_64_16:RougeL	0.537368646	0.706556881	0.544386877

Classification report in term of confusion matrix for one of the NN model: NN_64_16 has been shown below:

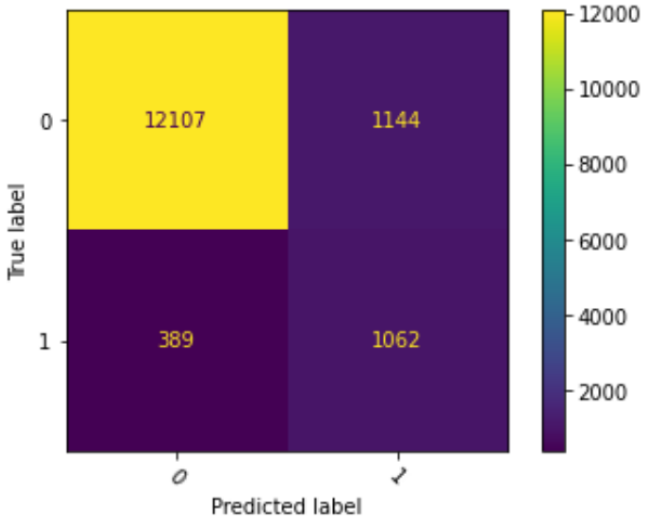


Figure 9: Confusion matrix for NN_64_16.

4) **LEAD3 and LR balanced as the baseline model:** Text summary prediction for LEAD3 and LR model with balanced weights as baseline model is obtained. Here, the LR model with balanced weight is trained with the sentence number as an additional feature to the sentence embedding and article mean. The ROUGE performance metrics for the predicted summary are given below:

S. N.	ModelID	recall	precision	F1-Score
14	LEAD3:Rouge1	0.576468613	0.764285021	0.583710494
15	LEAD3:RougeL	0.560056588	0.741543568	0.568121534
16	LR_bal:Rouge1	0.478130853	0.687714714	0.499510186
17	LR_bal:RougeL	0.448988922	0.647302445	0.470664301

5) **LSTM models:** LSTM models under different cases are trained with the data having sentence number and the article mean as additional features [6]. We consider default LSTM and bidirectional LSTM, both under two cases of number of hidden nodes. Therefore, the four cases of LSTM model are summarised as follows:

```
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 lstm_1 (LSTM)                (None, None, 25)          79400

 time_distributed_1 (TimeDis  (None, None, 1)           26
 tributed)

=================================================================
Total params: 79,426
Trainable params: 79,426
Non-trainable params: 0
```

Figure 10: LSTM with 25 hidden nodes: LSTM_25.

```
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 lstm_2 (LSTM)                (None, None, 50)          163800

 time_distributed_2 (TimeDis  (None, None, 1)           51
 tributed)

=================================================================
Total params: 163,851
Trainable params: 163,851
Non-trainable params: 0
_________________________________________________________________
```

Figure 11: LSTM with 50 hidden nodes: LSTM_50.

```
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 bidirectional_1 (Bidirectio  (None, None, 50)          158800
 nal)

 time_distributed_3 (TimeDis  (None, None, 1)           51
 tributed)

=================================================================
Total params: 158,851
Trainable params: 158,851
Non-trainable params: 0
_________________________________________________________________
```

Figure 12: Bidirectional LSTM with 25 nodes: LSTM_bi_25.

```
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 bidirectional_2 (Bidirectio  (None, None, 100)        327600
 nal)

 time_distributed_4 (TimeDis  (None, None, 1)          101
 tributed)

=================================================================
Total params: 327,701
Trainable params: 327,701
Non-trainable params: 0
_________________________________________________________________
```

Figure 13: Bidirectional LSTM with 50 nodes: LSTM_bi_50.

All of the above LSTM models are trained with Adam optimiser to minimise cross-entropy loss. After training, the model is tested with test samples. The result summaries for each model are presented in the table below.

S. N.	ModelID	recall	precision	F1-Score
1	LSTM_25:Rouge1	0.54841	0.761568	0.568592
2	LSTM_25:RougeL	0.530603	0.734615	0.550832
3	LSTM_50:Rouge1	0.554123	0.762394	0.573243
4	LSTM_50:RougeL	0.536463	0.735191	0.555563
5	LSTM_bi_25:Rouge1	0.552088	0.760683	0.570516
6	LSTM_bi_25:RougeL	0.533998	0.73399	0.552702
7	LSTM_bi_50:Rouge1	0.555638	0.764624	0.575063
8	LSTM_bi_50:RougeL	0.537943	0.737781	0.557474

Recall-Oriented Understudy for Gisting Evaluation:

Based the predicted summary, Rouge score for recall, precision and F-score has been calculated and all the results are tabulated as presented along with the corresponding algorithm description. Now we present the actual comparison of the predicted summary with the original summary.

S.N.	Summary by	summary	label
0	Original	The battle lines have been drawn between Pfizer's owners and managers, who will assemble on Thursday at the annual shareholder meeting in Lincoln, Neb., at the Cornhusker Marriott hotel. "Shareholders should withhold their votes for the four nominees of the Pfizer board of directors who are members of the board's compensation committee. This would be a first step on a long road to restore director accountability to owners."	0
1	Original	I like quiet." So, for those keeping score at home Mari isn't suffering writer's cramp. Dole is cramping her writing style.	1
2	LR_Default	By MATT SCHWARTZ in Houston and WENDELL JAMIESON in New York Daily News WritersSaturday, October 14th 1995, 4:22AMBleeding from a massive chest wound, Tejano star Selena cried, "Help me! Help me! Shaken witnesses yesterday told a spellbound Houston courtroom how the blood-covered, mortally wounded 23-year-old Hispanic singing sensation burst into the lobby of the Corpus Christi Days Inn last March 31.	0
3	LR_Default	By HOLLY RAMER, Associated PressCONCORD, N.H. -- A sick American engineer who was successfully evacuated from the South Pole to New Zealand is awaiting the results of medical tests after having what doctors believed was a stroke in August. She added, "So nice to see green and smell freshly cut grass, flowers, birds chirping, insects, etc., since it's now been over a year on the flat polar plateau of just ice and snow." She asked for an emergency evacuation after having what doctors believed was a stroke in August, but officials rejected her request because of bad weather, saying that sending a rescue plane was too dangerous and that her condition wasn't life-threatening.	1
4	Elastic Net	By MATT SCHWARTZ in Houston and WENDELL JAMIESON in New York Daily News WritersSaturday, October 14th 1995, 4:22AMBleeding from a massive chest wound, Tejano star Selena cried, "Help me! Help me! Shaken witnesses yesterday told a spellbound Houston courtroom how the blood-covered, mortally wounded 23-year-old Hispanic singing sensation burst into the lobby of the Corpus Christi Days Inn last March 31.	0
5	Elastic Net	By HOLLY RAMER, Associated PressCONCORD, N.H. -- A sick American engineer who was successfully evacuated from the South Pole to New Zealand is awaiting the results of medical tests after having what doctors believed was a stroke in August. She added, "So nice to see green and smell freshly cut grass, flowers, birds chirping, insects, etc., since it's now been over a year on the flat polar plateau of just ice and snow." She asked for an emergency evacuation after having what doctors believed was a stroke in August, but officials rejected her request because of bad weather, saying that sending a rescue plane was too dangerous and that her condition wasn't life-threatening.	1
6	Neural Net: NN_32_8	By MATT SCHWARTZ in Houston and WENDELL JAMIESON in New York Daily News WritersSaturday, October 14th 1995, 4:22AMBleeding from a massive chest wound, Tejano star Selena cried, "Help me! Help me! Shaken witnesses yesterday told a spellbound Houston courtroom how the blood-covered, mortally wounded 23-year-old Hispanic singing sensation burst into the lobby of the Corpus Christi Days Inn last March 31.	0
7	Neural Net: NN_32_8	By HOLLY RAMER, Associated PressCONCORD, N.H. -- A sick American engineer who was successfully evacuated from the South Pole to New Zealand is awaiting the results of medical tests after having what doctors believed was a stroke in August. She said results will be shared with doctors in the United States, "so everyone will be on the same page. ""Back at hotel now to chill out," Douceur wrote.	1
8	LSTM_25	By MATT SCHWARTZ in Houston and WENDELL JAMIESON in New York Daily News WritersSaturday, October 14th 1995, 4:22AMBleeding from a massive chest wound, Tejano star Selena cried, "Help me! Help me! Shaken witnesses yesterday told a spellbound Houston courtroom how the blood-covered, mortally wounded 23-year-old Hispanic singing sensation burst into the lobby of the Corpus Christi Days Inn last March 31.	0
9	LSTM_25	By HOLLY RAMER, Associated PressCONCORD, N.H. -- A sick American engineer who was successfully evacuated from the South Pole to New Zealand is awaiting the results of medical tests after having what doctors believed was a stroke in August. Renee-Nicole Douceur told The Associated Press in an email Tuesday that she had MRI and echocardiogram exams. She asked for an emergency evacuation after having what doctors believed was a stroke in August, but officials rejected her request because of bad weather, saying that sending a rescue plane was too dangerous and that her condition wasn't life-threatening.	1

Original, Class:0

The battle lines have been drawn between Pfizer's owners and managers, who will assemble on Thursday at the annual shareholder meeting in Lincoln, Neb., at the Cornhusker Marriott hotel. "Shareholders should withhold their votes for the four nominees of the Pfizer board of directors who are members of the board's compensation committee. This would be a first step on a long road to restore director accountability to owners."

Original, Class-1

I like quiet." So, for those keeping score at home Mari isn't suffering writer's cramp. Dole is cramping her writing style.

LR_Default, Class:0

By MATT SCHWARTZ in Houston and WENDELL JAMIESON in New York Daily News WritersSaturday, October 14th 1995, 4:22AMBleeding from a massive chest wound, Tejano star Selena cried, "Help me! Help me! Shaken witnesses yesterday told a spellbound Houston courtroom how the blood-covered, mortally wounded 23-year-old Hispanic singing sensation burst into the lobby of the Corpus Christi Days Inn last March 31.

LR_Default, Class:1

By HOLLY RAMER, Associated PressCONCORD, N.H. -- A sick American engineer who was successfully evacuated from the South Pole to New Zealand is awaiting the results of medical tests after having what doctors believed was a stroke in August. She added, "So nice to see green and smell freshly cut grass, flowers, birds chirping, insects, etc., since it's now been over

a year on the flat polar plateau of just ice and snow." She asked for an emergency evacuation after having what doctors believed was a stroke in August, but officials rejected her request because of bad weather, saying that sending a rescue plane was too dangerous and that her condition wasn't life-threatening.

Elastic Net, Class:0

By MATT SCHWARTZ in Houston and WENDELL JAMIESON in New York Daily News WritersSaturday, October 14th 1995, 4:22AMBleeding from a massive chest wound, Tejano star Selena cried, "Help me! Help me! Shaken witnesses yesterday told a spellbound Houston courtroom how the blood-covered, mortally wounded 23-year-old Hispanic singing sensation burst into the lobby of the Corpus Christi Days Inn last March 31.

Elastic Net, Class:1

By HOLLY RAMER, Associated PressCONCORD, N.H. -- A sick American engineer who was successfully evacuated from the South Pole to New Zealand is awaiting the results of medical tests after having what doctors believed was a stroke in August. She added, "So nice to see green and smell freshly cut grass, flowers, birds chirping, insects, etc., since it's now been over a year on the flat polar plateau of just ice and snow." She asked for an emergency evacuation after having what doctors believed was a stroke in August, but officials rejected her request because of bad weather, saying that sending a rescue plane was too dangerous and that her condition wasn't life-threatening.

Neural Net: NN_32_8, Class:0

By MATT SCHWARTZ in Houston and WENDELL JAMIESON in New York Daily News WritersSaturday, October 14th 1995, 4:22AMBleeding from a massive chest wound, Tejano star Selena cried, "Help me! Help me! Shaken witnesses yesterday told a spellbound Houston courtroom how the blood-covered, mortally wounded 23-year-old Hispanic singing sensation burst into the lobby of the Corpus Christi Days Inn last March 31.

Neural Net: NN_32_8, Class:1

By HOLLY RAMER, Associated PressCONCORD, N.H. -- A sick American engineer who was successfully evacuated from the South Pole to New Zealand is awaiting the results of medical tests after having what doctors believed was a stroke in August. She said results will be shared with doctors in the United States, "so everyone will be on the same page. ""Back at hotel now to chill out," Douceur wrote.

LSTM_25, Class:0

By MATT SCHWARTZ in Houston and WENDELL JAMIESON in New York Daily News WritersSaturday, October 14th 1995, 4:22AMBleeding from a massive chest wound, Tejano star Selena cried, "Help me! Help me! Shaken witnesses yesterday told a spellbound Houston courtroom how the blood-covered, mortally wounded 23-year-old Hispanic singing sensation burst into the lobby of the Corpus Christi Days Inn last March 31.

LSTM_25, Class:1

By HOLLY RAMER, Associated PressCONCORD, N.H. -- A sick American engineer who was successfully evacuated from the South Pole to New Zealand is awaiting the results of medical tests after having what doctors believed was a stroke in August. Renee-Nicole Douceur told The Associated Press in an email Tuesday that she had MRI and echocardiogram exams. She asked for an emergency evacuation after having what doctors believed was a stroke in August, but officials rejected her request because of bad weather, saying that sending a rescue plane was too dangerous and that her condition wasn't life-threatening.

Conclusions

This project has discussed the application of BERT embeddings for the text summarization for the news articles. The summaries from the news article were labelled by '0' or '1' based on the cosine similarity. Then, the embeddings from the train and test data were processed to obtained the domain features and document level features with doc labels. Therefore, train and test dataset were created by document level data split. With these train test data samples, various machine learning models were trained and tested for the text summary prediction. The prediction performance was analysed using the Rouge performance matrix in term of recall, precision and F-score. Also, the predicted summary was compared with the original summary for both the classes represented by '0' and '1'.

References

[1.] Available [online]: https://medium.com/luisfredgs/automatic-text-summarization-with-machine-learning-a n-overview-68ded5717a25.

[2.] Available [online]: https://towardsdatascience.com/nlp-extract-contextualized-word-embeddings-from-bert-keras-tf-67ef29f60a7b.

[3.] N. Andhale and L. A. Bewoor, "An overview of Text Summarization techniques," 2016 International Conference on Computing Communication Control and automation (ICCUBEA), 2016, pp. 1-7, doi: 10.1109/ICCUBEA.2016.7860024.

[4.] R. Boorugu and G. Ramesh, "A Survey on NLP based Text Summarization for Summarizing Product Reviews," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 352-356, doi: 10.1109/ICIRCA48905.2020.9183355.

[5.] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 International Conference on Data Science and Communication (IconDSC), 2019, pp. 1-3, doi: 10.1109/IconDSC.2019.8817040.

[6.] P. Janjanam and C. P. Reddy, "Text Summarization: An Essential Study," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 2019, pp. 1-6, doi: 10.1109/ICCIDS.2019.8862030.