CS 6120: Natural Language Processing
Project Report

Text Summarization

By: Simran Girdhar Kodwani, Aishwarya Mathur

I. Introduction

Natural Language Processing is a branch of artificial intelligence which gives machines the ability to understand human language. Text Summarization is a technique of reducing the number of words in a corpus without changing its meaning. There are various applications of text summarization which include converting customer reviews into meaningful smaller versions to take necessary actions, converting newspaper articles into small snippets which can be used in apps like inshorts, creating small summary reports from business meetings.

There are two types of text summarizations which are **extractive text summarization** and **abstractive text summarization**. Extractive text summarization selects the most important subset of sentences from the input text. This method extracts the most important information from the original text. This method is incapable of text generation and therefore the output is always a subset of the original text. Abstractive text summarization understands the core of the input text and produces a summary based on this understanding. This method is slightly more complex than extractive summarization as it tends to learn patterns in the input text. The summary this method produces doesn't necessarily have sentences from the input text corpus.
While abstractive summarization sounds more promising for its deeper understanding of the text corpus than extractive method, the extractive method as its own advantages such as these are easier to implement and need no prior training.

The model will be evaluated using various visualizations BLEU score and ROUGE metrics. BLEU (BiLingual Evaluation Understanding) is a metric for automatically evaluating machine-translated text. The BLEU score is a number between zero and one that measures similarity of machine-translated text to a set of high-quality reference translations. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics used for evaluating automatic summarization of texts as well as machine translations.

II. Libraries and Tools

The project will be implemented using Python language and Jupyter notebook. Python libraries such as Numpy, Pandas, Counter, SpaCy, NLTK, Keras, Genism, Tensorflow and BART will be used.

III. Dataset

1. Extractive Text Summarisation

The dataset used for extractive summarization is **BBC News Summary** dataset which contains news articles segregated into 5 categories namely, 'Business', 'Entertainment', 'Politics', 'Sport' and 'Tech' and the summaries of each article. The dataset contains a

total of 2,225 articles and summaries. It has 3 columns which are: title of article, article, and the summary.

	title	article	summary
0	Ad sales boost Time Warner profit	[Quarterly profits at US media giant TimeWarne...	TimeWarner said fourth quarter sales rose 2% t...
1	Dollar gains on Greenspan speech	[The dollar has hit its highest level against ...	The dollar has hit its highest level against t...
2	Yukos unit buyer faces loan claim	[The owners of embattled Russian oil giant Yuk...	Yukos' owner Menatep Group says it will ask Ro...
3	High fuel prices hit BA's profits	[British Airways has blamed high fuel prices f...	Rod Eddington, BA's chief executive, said the ...

2. Abstractive Text Summarization

The dataset used for abstractive text summarization is **CNN DailyMail News Text Summarization** dataset. This dataset is from Kaggle. It has 3 columns which are: id, article, highlights. The id column is a character column, article column has text data which is used as an input to the text summarization models and highlights column is a short summary of the articles which we will use to evaluate the performance of the model.

	id	article	highlights
0	0001d1afc246a7964130f43ae940af6bc6c57f01	By . Associated Press . PUBLISHED: . 14:11 EST...	Bishop John Folda, of North Dakota, is taking ...
1	0002095e55fcbd3a2f366d9bf92a95433dc305ef	(CNN) -- Ralph Mata was an internal affairs li...	Criminal complaint: Cop used his role to help ...
2	00027e965c8264c35cc1bc55556db388da82b07f	A drunk driver who killed a young woman in a h...	Craig Eccleston-Todd, 27, had drunk at least t...
3	0002c17436637c4fe1837c935c04de47adb18e9a	(CNN) -- With a breezy sweep of his pen Presid...	Nina dos Santos says Europe must be ready to a...

IV. Methods

1. Extractive Text Summarization

The extractive text summarising approach entails extracting essential words from a source material and combining them to create a summary. The output of extractive summarization can be thought of as a subset of the input text that communicates the document's main idea and emphasizes its key points. It is a straightforward model because it employs some method for rating the sentences in any text before ranking them and producing the results. This method works by identifying important text passages, removing them, and then piecing them back together to produce a condensed version. They therefore only use phrase extraction from the source text. For extractive summarization, TextRank algorithm is used which is a graph-based ranking model for text processing, based on *Google's PageRank* algorithm, that finds the most relevant sentences in a text.

For the project, the extractive summarization was performed on the BBC News Summary and the input news article category for the extractive summarization is taken from the user which includes Business, Entertainment, Politics, Sports, and Tech. After choosing the category, the article is split into sentences and the sentences were cleaned by performing methods like tokenization, conversion to lowercase, removing stop words and punctuation. The word vectors were obtained using GloVe global vectors and word embeddings were created. The first vectors (each of size 100 elements) for the constituent words in a sentence and were fetched to get the vector for the sentence. The sentence vectors were formed using the word embeddings and the sentences were compared with each other to form a similarity matrix using cosine similarity. Similarity scores between sentences were used to rank the sentences in the article and extract the most important points from the article

to form a summary. The model was also implemented using Gensim library, which has an inbuild summarization method which takes the text and the ratio parameter which specifies the length of summary to be formed and gives the summary of the article.

The extractive summarization model was evaluated by using ROGUE score metrics which works by comparing the automatically produced summary using the model against the reference summaries present in the dataset. It is calculated using the number of overlapping words and the total words present in the reference summary. Rouge-1 score refers to the overlap of unigrams and Rouge-2 score refers to the bigrams between the summary and reference summary. Rouge-L score measures the longest matching sequence of words using LCS (Longest Common Subsequence).

2. Abstractive Text Summarization

Transformers provide APIs and tools to download and train state of the art pretrained models. This reduces computational costs, carbon footprints and time to train a model from scratch. These models support various tasks such as text classification, named entity recognition, language modelling, summarization, image classification, object detection, audio speech recognition and video classification.

For the project, we used Hugging Face's pretrained transformer model. *Transformers* is a python library dedicated to supporting Transformer-based architecture and facilitating the distribution of pretrained models. At its core, the library is an implementation of the Transformer which is designed for both research and production.

The input text (the articles column from the dataset) was tokenized and then stop words and punctuations from the tokens were removed. On preliminary analysis, we noticed that words like *'by', 'associated', 'press', 'published' and 'updated'* appeared very frequently in the corpus, so we decided to drop these words too. The words in the corpus were lemmatized and the input was converted to a string. The input was thus pre-processed and ready to be fed into the model.

A function to generate summary was constructed. It counts the number of occurrences of each word in the input text, gets the maximum number of occurrences *say d* and divides this number *d* with the value of occurrence of each word to get a score. These scores are then used to get a sentence score for each sentence in the text. Then a *summarizer function* from the transformer library is called upon this text to generate a summary. The function also calculates the ROGUE score for the generated summary.

V. Results and Discussion

1. Extractive Text Summarization

Example of a news article from the dataset,

```
df['article'][2]
```

```
['The owners of embattled Russian oil giant Yukos are to ask the buyer of its former production unit to pay back a $900m (Â£479
m) loan.',
 '',
 'State-owned Rosneft bought the Yugansk unit for $9.3bn in a sale forced by Russia to part settle a $27.5bn tax claim against
Yukos. Yukos\' owner Menatep Group says it will ask Rosneft to repay a loan that Yugansk had secured on its assets. Rosneft alr
eady faces a similar $540m repayment demand from foreign banks. Legal experts said Rosneft\'s purchase of Yugansk would include
such obligations. "The pledged assets are with Rosneft, so it will have to pay real money to the creditors to avoid seizure of
Yugansk assets," said Moscow-based US lawyer Jamie Firestone, who is not connected to the case. Menatep Group\'s managing direc
tor Tim Osborne told the Reuters news agency: "If they default, we will fight them where the rule of law exists under the inter
national arbitration clauses of the credit."',
 '',
 'Rosneft officials were unavailable for comment. But the company has said it intends to take action against Menatep to recover
some of the tax claims and debts owed by Yugansk. Yukos had filed for bankruptcy protection in a US court in an attempt to prev
ent the forced sale of its main production arm. The sale went ahead in December and Yugansk was sold to a little-known shell co
mpany which in turn was bought by Rosneft. Yukos claims its downfall was punishment for the political ambitions of its founder
Mikhail Khodorkovsky and has vowed to sue any participant in the sale.']
```

Summary of the article,

```
df['summary'][2]
```

```
'Yukos\' owner Menatep Group says it will ask Rosneft to repay a loan that Yugansk had secured on its assets.State-owned Rosnef
t bought the Yugansk unit for $9.3bn in a sale forced by Russia to part settle a $27.5bn tax claim against Yukos.The sale went
ahead in December and Yugansk was sold to a little-known shell company which in turn was bought by Rosneft.But the company has
said it intends to take action against Menatep to recover some of the tax claims and debts owed by Yugansk."The pledged assets
are with Rosneft, so it will have to pay real money to the creditors to avoid seizure of Yugansk assets," said Moscow-based US
lawyer Jamie Firestone, who is not connected to the case.'
```

The accuracy and ROGUE score of extractive summarization model,

```
Accuracy:  0.54
```

```
Extracted Summary:
 Jamie Foxx and Hilary Swank have won the Screen Actors Guild Awards for best male and female film actors, boosting their Oscar
s hopes this month.Foxx's portrayal of late soul-singer Ray Charles in Ray had already earned him a prestigious Golden Globe aw
ard.Modest wine country comedy Sideways knocked out favourites Million Dollar Baby and The Aviator by taking the top prize for
best cast performance.Veteran actor Morgan Freeman took the best supporting actor award for playing a prize-fighter turned gym
manager in Million Dollar Baby."Thank you for Ray Charles for just living so complex and so interesting, and making us all just
come together," said Foxx, accepting his award in Los Angeles on Saturday.He also praised the film director: "Thank you for Tay
lor Hackford for taking a chance with an African-American film." Swank, too, was full of praise for her director and co-star Cl
int Eastwood.Both Foxx and Swank are now considered to be among the favourites to get Oscars - the Hollywood's ultimate prize.

Original Summary:
 Jamie Foxx and Hilary Swank have won the Screen Actors Guild Awards for best male and female film actors, boosting their Oscar
s hopes this month.Swank triumphed for playing a gutsy female boxer in Million Dollar Baby.Both Foxx and Swank are now consider
ed to be among the favourites to get Oscars - the Hollywood's ultimate prize.Swank, too, was full of praise for her director an
d co-star Clint Eastwood.Modest wine country comedy Sideways knocked out favourites Million Dollar Baby and The Aviator by taki
ng the top prize for best cast performance."I bow down to you," Swank said to the 74-year-old Eastwood.Veteran actor Morgan Fre
eman took the best supporting actor award for playing a prize-fighter turned gym manager in Million Dollar Baby.He also praised
the film director: "Thank you for Taylor Hackford for taking a chance with an African-American film.

Article summary: rouge1: 0.83 | rouge2: 0.75 | rougeL: 0.75
Average rouge: 0.8
```

The output of the model which uses TextRank algorithm using gensim,

```
Extracted Summary
 Modest wine country comedy Sideways knocked out favourites Million Dollar Baby and The Aviator by taking the top prize for bes
t cast performance.", '', 'The Screen Actors Guild (SAG) represents US film and TV actors.
Veteran actor Morgan Freeman took the best supporting actor award for playing a prizefighter turned gym manager in Million Doll
ar Baby.', '', '"Thank you for Ray Charles for just living so complex and so interesting, and making us all just come togethe
r," said Foxx, accepting his award in Los Angeles on Saturday.', '', 'He also praised the film director: "Thank you for Taylor
Hackford for taking a chance with an AfricanAmerican film.

Original Summary
 Jamie Foxx and Hilary Swank have won the Screen Actors Guild Awards for best male and female film actors, boosting their Oscar
s hopes this month.Swank triumphed for playing a gutsy female boxer in Million Dollar Baby.Both Foxx and Swank are now consider
ed to be among the favourites to get Oscars - the Hollywood's ultimate prize.Swank, too, was full of praise for her director an
d co-star Clint Eastwood.Modest wine country comedy Sideways knocked out favourites Million Dollar Baby and The Aviator by taki
ng the top prize for best cast performance."I bow down to you," Swank said to the 74-year-old Eastwood.Veteran actor Morgan Fre
eman took the best supporting actor award for playing a prize-fighter turned gym manager in Million Dollar Baby.He also praised
the film director: "Thank you for Taylor Hackford for taking a chance with an African-American film.

Article Summary:  rouge1: 0.56 | rouge2: 0.46 | rougeL: 0.46
Average rouge: 0.53
```

The ROUGE score of extractive summarization model is greater than the score by gensim summarizer for the same article.

2. Abstractive Text Summarization

The following is an example of an article from the dataset,

```
corpus[0]
```

```
". . . . . . . the bishop fargo catholic diocese north dakota exposed potentially hundred church member fargo grand f
ork jamestown hepatitis a virus late september early . the state health department issued advisory exposure anyone at
tended five church took communion . bishop john folda pictured fargo catholic diocese north dakota exposed potentiall
y hundred church member fargo grand fork jamestown hepatitis a . state immunization program manager molly howell say
risk low official feel 's important alert people possible exposure . the diocese announced monday bishop john folda t
aking time diagnosed hepatitis a . the diocese say contracted infection contaminated food attending conference newly
ordained bishop italy last month . symptom hepatitis a include fever tiredness loss appetite nausea abdominal discomf
ort . fargo catholic diocese north dakota pictured bishop located ."
```

It’s generated summary and corresponding ROGUE scores are,

```
summarize(corpus[0], df['highlights'][0])
```

```
Your max_length is set to 10, but you input_length is only 3. You might consider decreasing max_length manually, e.g.
summarizer('...', max_length=1)
```

```
 CNN.com will feature iRep  The bishop fargo catholic di  The state health department issued advisory exposure  Fargo
catholic diocese north  Bishop John folda diagnosed with hepatitis  symptom hepatitis a include fever tiredness  Farg
o catholic diocese north d
{'rouge1_fmeasure': tensor(0.1944),
 'rouge1_precision': tensor(0.1842),
 'rouge1_recall': tensor(0.2059),
 'rouge2_fmeasure': tensor(0.0571),
 'rouge2_precision': tensor(0.0541),
 'rouge2_recall': tensor(0.0606),
 'rougeL_fmeasure': tensor(0.1389),
 'rougeL_precision': tensor(0.1316),
 'rougeL_recall': tensor(0.1471),
 'rougeLsum_fmeasure': tensor(0.1667),
 'rougeLsum_precision': tensor(0.1579),
 'rougeLsum_recall': tensor(0.1765)}
```

VI. Future Scope

The objective of the upcoming research is to develop a powerful, domain- and language-independent text summarization that is effective with multiple documents. In a similar vein, the summary's quality evaluation, which is performed manually by knowledgeable assessors, is very individualized. There are specific criteria for evaluating quality, like grammar and coherence, but the outcomes vary when two different methods are used to assess the same article. The hyperparameters of the transformer model can be tuned further to achieve higher ROGUE scores. Furthermore, instead of using a transformer model, a sequence-to-sequence model can be used which will train an encoder and decoder model instead of using a pre-trained model.

VII. References

1. “CNN-DailyMail News Text Summarization.” Kaggle, 23 Oct. 2021, www.kaggle.com/datasets/gowrishankarp/newspaper-text-summarization-cnn-dailymail

2. Pietro, M. di. (2022, March 19). Text Summarization with NLP: TextRank vs Seq2Seq vs BART. Medium. https://towardsdatascience.com/text-summarization-with-nlp-textrank-vs-seq2seq-vs-bart-474943efeb09

3. Saxena, S. (2021, December 15). Text Summarization in Python using Extractive method (including end-to-end implementation). Medium. https://medium.com/analytics-vidhya/text-summarization-in-python-using-extractive-method-including-end-to-end-implementation-2688b3fd1c8c

4. Cachola, Isabel, et al. “TLDR: Extreme Summarization of Scientific Documents.” Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, 2020

5. “From Text Summarisation to Style-Specific Summarisation for Broadcast News.” Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2004, pp. 223–37

6. Richa Sharma, Prachi Sharma, “A Survey of Extractive Text Summarization”, International Journal of AdvancedResearchin Computer Science and Software Engineering, Volume 6, 2016.