News Summarization: Mandate-2

R Prasannavenkatesh
IMT2019063
IIIT Bangalore
prasannavenkatesh.ramkumar@iiitb.ac.in

M Mani Nandadeep
IMT2019051
IIIT Bangalore
mani.nandadeep@iiitb.ac.in

***Abstract*—The goal of this project is to build a text summarizer and use it to summarize the news articles. People are looking for shortcut methods to learn ideas in less time. Text summarizers are also helping people to decide whether a book, a research paper, or an article is worth reading or not. A text summarizer is a NLP based tool that wraps up text to a short length. It condenses a long article to main points. The need for text summarizers are increasing in everyday life since people want news quick and to the point rather than long and shabby walls of text. Social media sites like reddit also use a text summarizer called autotldr which is a bot that uses SMMRY to automatically summarize long reddit submissions.**

I. Introduction

Summarization is the task of condensing a piece of text to a shorter version, reducing the size of the initial text while at the same time preserving key informational elements and the meaning of content. Here, we perform text summarization on news articles corresponding to the dataset present at https://www.kaggle.com/sunnysai12345/news-summary.
The dataset consists of 4515 examples and contains Author name, Headlines, URL of Article, Short text, Complete Article. The news has been summarized from Inshorts and only scraped the news articles from Hindu, Indian times and Guardian in time period ranges from February to August 2017.

II. Preprocessing

It is critical to process data before using it for analysis or prediction. Text processing is a technique used in NLP to clean text and prepare it for model building. It is adaptable and contains noise in a variety of forms, such as emotions, punctuation, and text written in numerical or special character forms. The following are some of the preprocessing steps our current notebook is performing:

- Converting to lower case to ensure standard caseing system throught the process.
- Removing stop words from the text. Stopwords are frequently used words that are removed from the text because they add no value to the analysis. We are using nltk package and english stopwords to perform stopword removal. Once the process is completed, stop words that are present in the nltk library are removed from the tokenized text and the rest are stored.
- Punctuations and other symbols are removed using regex and replacing the matched pattern with an empty string.
- URLs are removed using regex and replacing the matched pattern with an empty string.
- Stemming is performed on the text where the words are stemmed or diminished to their root/base form. Snowball stemmer and Porter Stemmer algorithm both are implemented in the code using the functions provided by nltk. However, Porter stemmer algorithm is used because of it's better results. The Porter stemming algorithm is a process for removing the commoner morphological and inflexional endings from words in English.
- Lemmatization stems the word while ensuring that it retains its meaning. It has a pre-defined dictionary that saves the context. Lemmatization is implemented using the WordNetLemmatizer provided by nltk package. Wordnet is an large, freely and publicly available lexical database for the English language aiming to establish structured semantic relationships between words. It offers lemmatization capabilities as well and is one of the earliest and most commonly used lemmatizers.

III. EDA

Exploratory Data Analysis (EDA) is an approach for data analysis that employs a variety of techniques (mostly graphical). Here are some of the EDA implemented:

- Plotted density histograms for text and summary

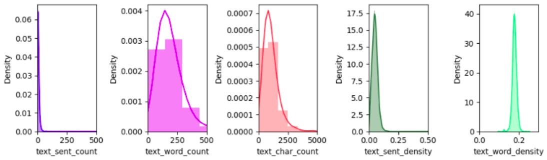

Fig. 1. Density histogram for text

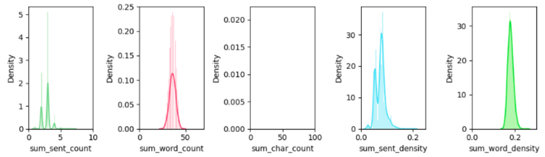

Fig. 2. Density histogram for summary

- Visualised Word Cloud

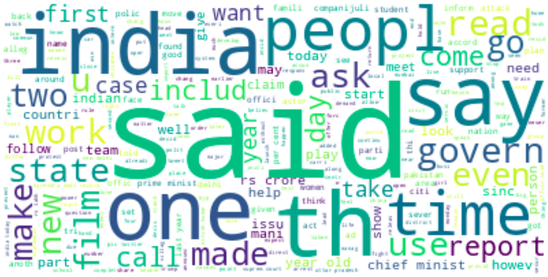

Fig. 3. Word cloud for text

IV. T5 Transformer

T5 is a new transformer model from Google that is trained in an end-to-end manner with text as input and modified text as output. It achieves state-of-the-art results on multiple NLP tasks like summarization, question answering, machine translation etc using a text-to-text transformer trained on a large text corpus.

We have implemented till T5 Tokenizer for this mandate submission.

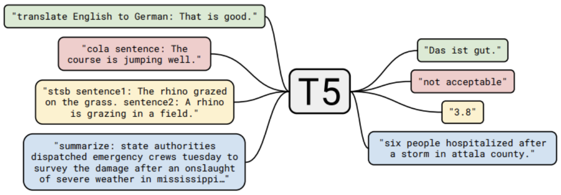
"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

Fig. 4. T5 Framework

V. Outcomes mapping to mandate contributions

- CO1: ****
- CO2: ****
- CO3: *
- CO4: -
- CO5: **
- C06: ***

VI. Work Split

- IMT2019063 - R Prasanna: Preprocessing, EDA
- IMT2019051 - Mani Nandadeep: tokenization

VII. References

- Text summarization EDA
- Documentation of T5
- Simple abstractive text summarization with pretrained T5