Social Media Comment Overview Generation

Group 3

Rathnapriya Gopalakrishnan (rgopala@iu.edu),
Akshya Ramesh (aksrame@iu.edu)

Introduction

Ever since the first social media website SixDegrees.com came into existence in 1997, thousands of other applications have evolved. These applications fall into a variety of categories such as media sharing applications (YouTube, Instagram, Snapchat, TikTok, etc.), networking applications (Facebook, Twitter, LinkedIn, etc.), discussion forums (Reddit, Quora, Stack Overflow, etc.), review sites (Yelp, Zomato, etc.) and shopping networks (Amazon, Flipkart, Etsy, Shien, etc.).**[1]** Approximately 4.6 billion people use social media, which is half of the world's population, who spend an average of 2 hours 27 mins per day. The most used social media platforms are Facebook, YouTube, and Instagram **[2].** A survey estimates that there are between 3.2 to 37.8 million influencers on the most popular social media platforms: YouTube, Instagram, and TikTok. **[3].** These platforms allow influencers to post images, videos, and written content in the form of posts, stories, short clips, etc. These applications allow viewers to provide feedback for the influencer's content in the form of likes, comments, and share features. Of these features, commenting is one that is most engaging as the viewer takes their time to provide feedback, show support, encouragement or make criticisms on the influencer's content. Big influencers with millions of followers usually get thousands of comments per post/image/video. Analyzing these comments and generating an overview of them might provide valuable information to the influencers. It could motivate them, help them explore suggestions, and improve their future content. It will also be useful for future viewers as it provides a short summary of the content from the perspective of past viewers. But the existing social media applications do not offer this functionality yet. This could lead to a huge loss of valuable data which is left unattended. We believe that enabling this functionally could have a huge impact on social media influencers as it makes them aware of what their viewer's opinions are and lets them know what viewers are looking for in their future content.

Problem Description

Many social media influencers have millions of followers who get thousands of comments on their content. None of these influencers have the time to go through each of these comments to understand how the viewers received their content. If the viewers feel that their voices are not being heard by the influencers, they might gradually lose interest in their content, and this leads to gradual follower dropout. This might affect the number of sponsorships the influencers get and lead to fewer earnings.

Our proposed model will provide a way for the influencers to get a gist of these comments. The model works by scanning through multiple comments for a particular content and generates a summarized paragraph which gives an overview of the comments. This overview will be helpful for both the influencers and the viewers, as influencers can see how their content was received and the viewers can get to know what to expect from the content based on previous viewers' experience.

Datasets

We are planning to use the following datasets.

1. **DART - Open-Domain Structured Data Record to Text Generation**

   This dataset contains 82191 records taken from various domains. Each sentence in the dataset is associated with its keywords in the form of triplets.

2. **WebNLG Dataset Summary**

   There are three versions of this dataset, of which we use webNLG_v2.1. It contains 42873 keyword-text pairs. The data has already been preprocessed and cleaned and made available in both JSON and XML formats.

3. **Trending YouTube Video Statistics and Comments**

   This dataset contains 806,932 YouTube comments trending from different categories collected each day along with the video statistics. Multiple comments for a single video are recorded in this dataset

4. **New York Times Comments**

   This dataset contains comments from the New York Times news articles recorded in Jan-May 2017 and Jan-April 2018. The dataset has a total of 2 million comments along with the articles, featuring various categories in a different file. We are planning to use this dataset for testing our fine-tuned model.

5. **Yelp Reviews Dataset**

   Yelp is a website that publishes user reviews for establishments such as restaurants, bars, cafes, spas etc. This dataset was obtained from Kaggle, and it contains 10000 yelp reviews. Multiple reviews for each establishment are recorded.

Proposed Methodology

The steps involved in our project are as follows

1. Collecting the data from multiple sources (DART, WebNLG, YouTube Comments and Yelp Review dataset)
2. Extracting triplets from the data and storing them along with the original sentences
3. Fine tuning an existing text summarization model with generated triplet dataset
4. Evaluating the model on the New York Times comment dataset

Data Collection

For this project, we use five data sources. The WebNLG, DART, YouTube Comments, and Yelp Review datasets are used in the training phase and the New York Times comment dataset is used for testing and evaluation of the model. Both the WebNLG and DART datasets contain triplets for various sentences which are used in this project. For the YouTube Comments dataset, we utilize comments stored in the ‘comment_text‘ column, for the Yelp Reviews dataset we use reviews stored in the ‘text’ column, and in the New York Times comment dataset we use the comments stored in the ‘commentBody’ column.

Data Preprocessing

The collected data is cleaned by removing missing values and removing duplicate values. This cleaned dataset is then used for extracting triplets from each sentence, which can be done using the Stanford's Open Information Extraction (Open IE).[10] This identifies the triplets in the sentence using the Treebank parser tree through the Stanford parser. Some triplet variations are Subject-Predicate-Object and Subject-Verb-Object. Using Part of Speech Tagging (POS), it is also possible to extract triplets which consider Adjectives and Adverbs.

The DART, WebNLG already has the triplets generated, so our aim is to generate these triplets for only YouTube video comments and Yelp review dataset and map it with the original sentence. These triplets will be fed to the pre-trained text summarization model along with the original sentences for supervised learning.

Text Summarization Model

There are many text summarization transformer models in the NLP realm, such as GPT-2, BART, XLM, T5, etc. The T5 model can be easily fine-tuned to various downstream tasks and has also given state of art performance on many benchmarks such as GLUE, SQUAD, etc. Thus, we decided to use the T5 model for text summarization using triplets.

T5 Model

The T5 is a Text-to-Text Transfer Transformer model introduced by Google [9] in early 2020. The T5 model uses the technique of Transfer Learning. This involves pretraining a model on large unsupervised data and later fine-tuning it on smaller supervised data for downstream tasks. The T5 model can perform various NLP tasks such as translation, summarization, text similarity, classification, regression, etc.

Fine Tuning with WebNLG

We used the technique explored by Mattew Alexander [11] to fine tune the T5-small model with the webNLG data for text summarization through triplets. We got the below text summarization results when provided the keywords as input.

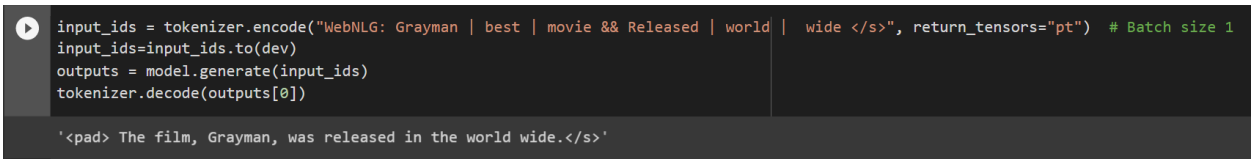

Fig 1: Original Sentence: Grayman is the best movie, and it is released worldwide.

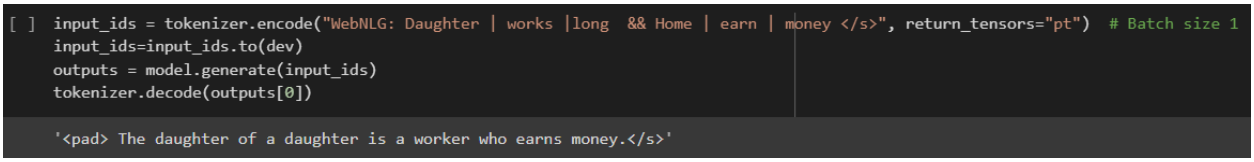

Fig 2: Original Sentence: My daughter works long hours. So, the home can earn more money.

```
[17] input_ids = tokenizer.encode("WebNLG: I | loved | video && You | inspire |  everyone </s>", return_tensors="pt")  # Batch size 1
     input_ids=input_ids.to(dev)
     outputs = model.generate(input_ids)
     tokenizer.decode(outputs[0])

     '<pad> I was a video artist and inspires you.</s>'
```

Fig 3: Original Sentence: I loved all your videos. You inspire everyone to do great things.

As seen from the above outputs, we can note that, for some inputs (Fig 1), the model generated cohesive sentences that captured the partial essence of the input data. Whereas for others (Fig 2 and Fig 3), it failed to generate a comprehensive and meaningful paragraph, and so these summarizations do not reflect the meaning of original sentences.

In order to improve the summarization capability of the T5 model, we suggest fine tuning it with additional datasets such as DART, YouTube comment and Yelp review dataset.

The DART dataset already contains triplets for each sentence which we use to fine tune the T5 model. But there are not many other triplet datasets available on the internet. So, we propose a way to convert comments from YouTube and Yelp dataset into triplets using Standford Open IE to further fine tune the model.

Evaluation Plan

There are many evaluation metrics for text summarizations, such as BLEU scores, ROUGE N, ROUGE L, ROUGE S, BERTScore etc. Among these we will use **ROUGE N** [12] and **BERTScore** [13]. For the evaluation of our model, we will feed our test dataset, which is the triplets obtained from the New York Times comment data. The results from the fine-tuned model are then evaluated using Recall-Oriented Understudy for Gisting Evaluation (ROUGE). The ROUGE evaluation compares the summarized results from the model with the reference summary. We will use ROUGE N with n=1 (unigram) which checks to see if the unigrams in the summarized results appear in the unigrams in the summary. ROUGE Recall tells us how many unigrams in the reference summary appear in the summarized result. ROUGE Precision tells us how many unigrams in the summarized results contain the unigrams in the reference summary. In order to obtain a compromise between the ROUGE Precision and ROUGE Recall, we use F1 Score which is a harmonic mean of those two. This ROUGE F1 Score is our evaluation metric.

The main disadvantage of ROUGE evaluation is that the words need to be exactly the same. It does not consider words that are semantically correct. This means that if “excellent” is used instead of “outstanding”, ROUGE will not accept it.

To overcome this, we will use BERTScore, which allows for semantically matching words to appear in place of the original word. BERTScore uses contextual word embeddings and cosine similarity between reference summary and the model generated summary to achieve this.

Expected Outcomes

From the above approach, we obtain a model summarizing comments and reviews for influencer’s contents. This model can enable them to get a shortened version of how their content has been received by its viewers. From the evaluation results, we can choose to improve our model by using

fine tuning it with larger datasets and with a larger T5 size model such as t5-base, t5-large, t5-3b or t5-11b.

The model can be further improved by identifying and removing irrelevant and spamming comments made on the influencer's comment section. This project can be extended to summarize other language comments too by translating them to English.

References

1. https://www.digitalvidya.com/blog/types-of-social-media/
2. https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/
3. https://earthweb.com/how-many-influencers-are-there/#:~:text=Our%20data%20reveals%20that%20there,Instagram%2C%20YouTube%2C%20and%20TikTok.
4. https://github.com/Yale-LILY/dart
5. https://gitlab.com/shimorina/webnlg-dataset
6. https://www.kaggle.com/datasets/aashita/nyt-comments
7. https://www.kaggle.com/datasets/datasnaek/youtube?select=GBcomments.csv
8. https://www.kaggle.com/datasets/omkarsabnis/yelp-reviews-dataset
9. https://arxiv.org/abs/1910.10683
10. https://github.com/philipperemy/Stanford-OpenIE-Python
11. https://towardsdatascience.com/data-to-text-generation-with-t5-building-a-simple-yet-advanced-nlg-model-b5cce5a6df45
12. https://towardsdatascience.com/the-ultimate-performance-metric-in-nlp-111df6c64460
13. https://towardsdatascience.com/bertscore-evaluating-text-generation-with-bert-beb7b3431300