Generating Effective Headlines for Social Media Posts

Raymond Hung, Jack Forlines, Dipika Kumar
DATASCI 266: Natural Language Processing with Deep Learning
UC Berkeley School of Information

Abstract

This study utilizes pre-trained LLMs to generate headlines from social media posts, emphasizing accuracy and engagement. By employing models like PEGASUS, T5, and BART on Reddit data, we aim to preserve semantic content while enhancing summarization. Evaluation metrics, including ROUGE, BLEU, and semantic similarity scores, alongside human feedback, highlight the T5 model's superior performance. Despite computational constraints and model-specific challenges, our research underscores the importance of innovative approaches to evaluation of headline generation from user-generated content.

1 Introduction

This study seeks to leverage advanced pre-trained large language models (LLMs) to address a text summarization challenge. Specifically, we aim to create headlines from extensive social media posts on Reddit, focusing on capturing their core message, relevance, and semantic content. As digital consumers increasingly adapt to the format of online articles and social media posts, it is crucial for digital-native companies to maintain high content quality to enhance user engagement and retention. An effective headline should be succinct, engaging, and accurately reflect the main ideas of the text. Search engines such as Google analyze headlines to assess content relevance, which influences the PageRank of websites. Higher PageRank improves a site's visibility from an SEO perspective, leading to increased web traffic and click-through rates. For digital-native companies, maintaining high-quality content is vital, as it forms the backbone of their business model.

User-generated content, especially text, relies on its users to craft headlines that accurately reflect the content of their posts in a captivating way. However, these user-created titles may sometimes lack accuracy or appeal which highlights a significant opportunity for enhancement through recent advances in NLP and LLMs for text summarization. A practical application could involve using LLMs to either suggest titles or auto-correct existing ones based on the content, potentially resulting in titles that are more precise, concise, and engaging than those written by humans. Many companies have integrated text summarization into their products. For instance, Gmail features auto-generated suggested titles and eBay has incorporated automated title generation into its listing service to enhance the precision of user listings.

While the potential applications of LLMs in real-world scenarios are promising, challenges remain, particularly in the realm of summarization. Extractive summarization methods excel in maintaining the faithfulness of the generated content but often fall short in fluidity. Conversely, abstractive summarization offers more natural readability but carries a higher risk of factual inaccuracies, as the generation of new text may introduce errors not present in the original content. This is especially relevant in the context of social media, where preserving factual integrity is essential. Additionally, the summarization of social media posts often necessitates eye-catching headlines to attract potential readers and enhance click-through rates.

As these summarization techniques gain wider adoption, it will be imperative for organizations to apply stringent evaluation practices to ensure that the generated outputs maintain a natural flow of content but also uphold factual integrity. Our research aims to explore the efficacy of text summarization to generate headlines through the application of pre-trained sequence-to-sequence (seq2seq) models, including PEGASUS, T5, and BART, applied to Reddit posts and their corresponding titles. We evaluate performance primarily through metrics such as ROUGE, BLEU, Angular Embedding Similarity, and Fréchet's Distance, complemented by qualitative feedback from human evaluators. .

2 Related Work

Headline Generation has been a significant area of focus and advancement in NLP in news and social media where studies have been conducted to generate more engaging and accurate headlines for users. Significant progress has been made in leveraging user reading histories to generate more personalized headlines (Cai et al., 2023). This has highlighted its potential to improve news recommendation systems with tailored content.

In a recent study, PEGASUS and BART models were fine-tuned on social media posts to improve headline appeal. The authors employ a disentanglement-based model to effectively balance the content and contextual attributes of generated headlines. This approach yielded significant enhancements in headline generation, as confirmed by ROUGE metrics and human evaluations (Zhang Yang, 2023). Many challenges remain in summarization particularly for social media content. Traditionally, abstractive summarization research has predominantly focused on news articles, which are more structured in contrast with the informal and varied nature of social media text. Research findings have highlighted the limitations of applying advanced summarization for social media. Furthermore, when assessing abstractive summaries, it has proven challenging to reach clear conclusions using ROUGE scores alone, thus warranting extensive manual evaluation (Syed et al., 2019).

ROUGE, a recall-oriented method for abstractive summarization evaluation, scores n-gram overlap (Lin, 2004). Yet, this focus may not gauge truthfulness or factual accuracy effectively (Matsumaru et al., 2020). Recent studies propose alternative metrics like Angular Embedding Similarity and Fréchet Distance to precisely measure semantic similarity (Moeed et al., 2020).

3 Methods

3.1 Data Collection and Cleansing

Our Reddit Dataset includes more than 500,000 posts from 19 data science and machine- learning related subreddits, sourced through the pushshift.io API and available on Kaggle. For our summarization task, we remove irrelevant columns, keeping only the variable-length text fields for each post and title. The dataset uses the body of each post as input and the corresponding user-generated titles as output for training and fine-tuning our models. These titles typically provide a concise and catchy summary of the post content. As part of data cleansing, entries with *NaN* values were eliminated. Text normalization was applied, including stripping HTML tags, URLs, special characters, and punctuation, resulting in enhanced dataset quality. However, upon further review, it was noted that posts containing code examples were not controlled for. Through preprocessing, our dataset was further refined to 274,209 pairs of titles and posts.

3.2 Design and Implementation All the models were trained on either V100 or L4 GPUs on Colab. Due to limited GPU resources, we were not able to use the complete population of our dataset. Therefore, we shuffled and exported 25,000 ($\tilde{9}$%) of our refined records as a subset for fine-tuning. We then applied an 80% train, 15% test, and 5% validation split to our pairs data for further processing and fine-tuning. Considering that PEGASUS (Zhang et al) is extensively trained and fine-tuned on news articles and widely employed for generating headlines for social media posts (Zhang & Yang, 2023), we selected the *HuggingFace* model checkpoint "Extreme Summarization (XSum)" as our baseline model. This decision was based on PEGASUS' demonstrated ability to generate concise and informative summaries, making it suitable as a starting point for headline generation. We included comparisons with other widely recognized transformer-based models such as T5 (Raffel et al), BART (Lewis et al), OPT (Zhang et al), and GPT-2 (Radford et al) to provide a comprehensive evaluation.

3.3 Fine Tuning We compare our baseline model (Pegasus-Xsum) against the other seq2seq and transformer models. All models were fine-tuned using the same pairs dataset, ensuring a fair comparison of their performance. Additionally, we attempted to fine-tune BART and OPT350m models; however, these attempts were unsuccessful due to resource constraints when investigating technical difficulties with the specific requirements for our headline generation task. We did not achieve generating results on the size of our test sample for these models. It is further worth noting that the GPT-2 model failed in producing satisfactory summarizations compared to other models, however we achieved some automated measurements. This outcome could be attributed to the GPT-2 model architecture not being well-suited for summarization tasks, or issues present with the implementation code used for fine-tuning the model. As noted previously, resource constraints impacted our ability to produce satisfactory results for the GPT-2, BART, and OPT350m models, although testing on smaller samples of test pairs for BART and OPT350m produced promising results confirming their potential for the task of headline generation. As the T5 model produced superior evaluation metrics at equal hyperparameters across all models, we further fine-tuned the T5 model. After experimenting with max_length, batch_size, and learning_rate hyperparameters, we found increasing the max_length from 32 to 64 resulted in a noticeable increase in both training and validation accuracy, as well as decrease in training and validation losses. We kept the batch_size at 16 as decreasing it to 8 resulted in a much lower training loss than validation, which indicated overfitting. We also experimented by reducing the learning_rate from .0005 to .00005 and increasing it to .001, and these did not impact the results. The positive impact of the max_length increase is likely attributed to the enhanced context in the embeddings.

3.4 Evaluation

ROUGE and BLEU scores were used as a starting point to evaluate the efficacy of generated headlines. While ROUGE and BLEU provide valuable insights into the quality of the generated headlines, we also explored embedding-based metrics such as Angular Embedding Similarity Scores (AES) (Moeed et al) to further analyze the semantic quality of translated text. We spot-checked models with sample outputs to uncover anything that quantitative metrics did not cover. For example, the GPT-2, OPT350m, and BART models were producing the same outputs as the input fed into it. These models required specific preprocessing and data generation stepss. We attempted to streamline data generation and preprocessing steps across our models to limit resource constraints, however this introduced obstacles when training GPT-2, BART, and OPT350m. For example, GPT-2's autoregressive nature, designed for generating text following from what was previously given, could have been underutilized if the input data did not effectively leverage its capabilities. Without a clear prompt to generate something new, GPT-2, in particular, tended to regurgitate the input. Further data preprocessing is required to improve GPT-2 to focus more on straight sequence processing to leverage the model's autoregressive nature. This case is also true for BART and OPT350m models, hence why results from these models are omitted from this report as we were unable to train these models appropriately on the full corpus of data.

CandTitle-RefTitle	rouge1	rouge2	rougeL
PEGASUS-Xsum	0.157997	0.043143	0.14421
T5	0.184472	0.051011	0.165052
GPT-2	0.069966	0.023475	0.053999

Table 1

CanTitle-RefTitle	rougeLsum	bleu
PEGASUS-Xsum	0.143903	0.019779
T5	0.165165	0.029681
GPT-2	0.053998	0.007463

Table 2

To supplement our automated evaluation metrics, we conducted a human evaluation study to assess the quality of the generated headlines from a reader's perspective. We randomly selected a subset of test pairs and their corresponding generated headlines from T5. We then created a brief survey that presented the original Reddit post along with the generated headlines, without revealing which model generated each headline. The survey was distributed to a group of random participants (N=10) who were asked to select the headline they considered the most effective in capturing the essence of the Reddit post. This human feedback helps to provide a more comprehensive assessment of the model's performance in generating high-quality, engaging headlines for the given dataset.

4 Results and Discussion

4.1 Automatic Evaluation BLEU and ROUGE

Using a subset of the original test pairs dataset, we produced the titles (cand_titles) for the enhanced T5 model using an increased max_length hyperparameter. Evaluating the candidates against the original titles using the same metrics as earlier, we found increased scores across the board (**Tables 3 and 4**).

Angular Embedding Similarity and Frechet Distance For a more comprehensive evaluation, we added Angular Embedding Similarity (AES) and Fréchet Distance (FD) to assess the semantic similarity among generated titles (cand_title), reference titles (ref_title), and original posts (ref_post).

CanTitle-RefTitle	rouge1	rouge2	rougeL
T5-MaxL64	0.239797	0.08335	0.219164

Table 3

CanTitle-RefTitle	rougeLsum	bleu
T5-MaxL64	0.220178	0.059214

Table 4

AES	CanTitle-RefPost	RefTitle-RefPost
PEGASUS-Xsum	0.63724	0.64045
T5	0.64847	0.64045
GPT-2	-	0.64045

Table 5

Although typically employed in computer vision tasks, these metrics are used to measure semantic similarity between the encodings of the original texts and the generated summaries with greater precision (Moeed et al., 2020). Angular Similarity, which is akin to Cosine Similarity, measures the normalized angle between two vectors, with values ranging from 0 to 1. A value close to 1 indicate that the embeddings share similar semantic meanings, while a value closer to 0 suggests they are dissimilar.

On the other hand, Fréchet Distance measures the similarity between two curves by considering the sequence and positioning of points along them. In the context of NLP, it serves as a quantitative metric for comparing semantic similarity between machined generated and human texts, with lower values indicating similarity.

4.2 Human Evaluation Before calculating AES and FD, the test samples for candidate titles, reference titles, and original posts were encoded into 512-dimensional vector embeddings using the Universal Sentence Encoder. This step ensures that the semantic content of each text is effectively captured for analysis. It was noted that both AES and FD effectively complemented ROUGE and BLEU metrics in evaluating the T5 model, shown in **Table 3**.

5 Conclusion

In our study utilizing pre-trained LLMs to generate headlines from Reddit posts, we trained five total models, w only three produced satisfactory results after preprocessing and tuning. The T5 model yielded the best results from our BLEU and ROUGE metrics. T5's superior performance might be attributed to it being a larger model and more adaptable to diverse tasks. Even though T5 performed the best among our automated results, further human evaluation is required for producing meaningful results. Our study also faced significant challenges due to compute resource constraints, which limited our ability to fully explore model capabilities and necessitated a focus on efficient training strategies. This work illustrates the practical difficulties of implementing advanced large language model refinement and underscores the need for ongoing innovation and resource management.

FD	CanTitle-RefPost	RefTitle-Ref Post
PEGASUS-Xsum	1.07554	1.06696
T5	1.04622	1.06696
GPT-2	0.01173	1.06696

Table 6

6 Appendices

References

- MediaHG: Rethinking Eye-catchy Features in Social Media Headline Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5766–5777, Singapore. Association for Computational Linguistics.

- Shahbaz Syed, Michael Völske, Nedim Lipka, Benno Stein, Hinrich Schütze, and Martin Potthast. 2019. Towards Summarization for Social Media - Results of the TL;DR Challenge. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 523–528, Tokyo, Japan. Association for Computational Linguistics.

- Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki. 2020. Improving Truthfulness of Headline Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1335–1346, Online. Association for Computational Linguistics.

- Abdul Moeed, Yang An, Gerhard Hagerer, and Georg Groh. 2020. Evaluation Metrics for Headline Generation Using Deep Pre-Trained Embeddings. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1796–1802, Marseille, France. European Language Resources Association.

- Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text*

Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.