Automatic Hashtag Generation for Social Media Posts Using Neural Text Generation Models

Satwik Goyal, Shubham Shetty, and **Mahidhar Kommineni**
College of Information and Computer Sciences
University of Massachusetts Amherst

Abstract

In 2021, about 82% of adults in the United States used social media platforms. Social Media is used for various purposes including information sharing, marketing, and social interaction using posts. Each post may contain hashtags that help reach out to a wider audience by making it easier to search through the content. Hashtags create instant communities with people of similar interests. Analysing social media posts can help in several tasks such as targeted marketing. We propose a framework to generate hashtags for a given social media post using massive pre-trained language models in order to increase a post's reach.

1 Introduction

1.1 Motivation

Social media has increasingly become an essential mode of communication as a platform for self-expression and dissemination of information. In 2021, about 82% of adults in the United States used social media platforms (Statista, 2022). Social media posts cover a wide range of topics ranging from personal, political, social topics, to promotional content. With the explosion in the amount of posts being created and shared over various websites such as Twitter, Instagram, and Facebook, social media becomes a challenging task to make your post stand out.

We define a social media post as a document shared on a social media platform (such as Twitter) which can consist of either text, images, and/or videos. The inherent structure of a social media post provides several challenges such as analyzing a social media post in isolation given how limited the data in the post could be. Additionally a post may be made in context to additional information such as comment chains, community posts, and replies which might not be included in the information extracted from the post.

Hashtags are very often used to reach out to a wider targeted audience, however, some hashtags based on their popularity are more optimal than others. Therefore, we propose a model that will help generate optimal hashtags for social media posts to reach to a wider targeted audience.

1.2 Task Description

Users may use #hashtags to associate their posts with certain topics and drive engagement. Hashtags can further be used for downstream social media analytical tasks such as search and targeted marketing. However, many social-media posts do not have any tagged hashtags (Wang et al., 2011). Our hypothesis is that hashtags can be generated through pre-trained language models such as GPT-3 using just the context present in the post. This data can be either text, image or video format.

Our main task involves building a pipeline which can accept a social media post of any format and pass it through our pre-trained language model to generate relevant hashtags. This pipeline will have different branches to handle text, image, and video data respectively. Generated hashtags should be relevant to the post and help drive engagement.

2 Background & Related Work

Our work aims at generating hashtags from text and image based social-media posts using massive pre-trained language models (PLMs).

2.1 Automatic Hashtag Generation

Automatic hashtag annotation has been historically seen as a key-phrase extraction and classification task. Some earlier proposed methods for hashtag annotation include predicting hashtags from a pre-defined list of words (the candidate list) (Zhang et al., 2017), or using topic models to generate hashtags (Wu et al., 2016). The issue with these approaches is that they only generate single phrases as hashtags, whereas actual hashtags may be longer phrases which actually reflect the topic being referenced to in the post.

2.2 Neural Text Generation

Another issue with treating hashtag generation as a key-phrase extraction task is that hashtags might not appear in either the target posts or the given candidate list. Recent approaches to hashtag generation explore neural text generation models in order to generate phrase level hashtags beyond the provided post or candidate list. Current state-of-the-art approach (Wang et al., 2019) proposes a sequence-to-sequence text generation framework to generate hashtags based on social media conversations as a supplementary data source. However, it is unrealistic to expect conversation chains to contain required information, and has a high cost of annotation.

Further work builds on the sequence-to-sequence model proposed by (Wang et al., 2019). One such approach uses a semantic-fragmentation-based selection mechanism in transformer architecture to generate hashtags over large datasets (Mao et al., 2021). Another proposed framework utilises a retrieval-augmented sequence-to-sequence architecture (Zheng et al., 2021) in order to generate hashtags with current and emerging events reflected.

An issue with these approaches are that they require either external information or large amounts of data to generate meaningful results.

2.3 Massive Pre-Trained Language Models for Text Generation

Our proposed approach will try to address these issues by utilizing massive pre-trained language models (PLMs) with autoencoders for text generation, such as GPT-3 (Brown et al., 2020), to generate hash-tags. We will also experiment with zero-shot approaches (Kumar et al., 2019) towards hashtag generation using these massive models.

GPT-3 (Brown et al., 2020) has shown near state-of-the-art performance in generative tasks such as summarization and news article generation in a few-shot or zero-shot setting (Tehranipour, 2020). We intend to perform similar experiments using zero-shot and few-shot approach, and compare hashtag generation performance against that of our other models.

2.4 Text Augmentation

Text augmentation involves augmenting the existing linguistic space to improve the performance of down-stream tasks without affecting the generalizability. There are two types of popular text augmentation techniques namely shallow and deep augmentation. In shallow augmentation technique, small noises in the form of words and phrases are added to the linguistic space to improve the performance of downstream tasks. EDA (Wei and Zou, 2019) employs shallow augmentation technique to improve the performance of text classification task. On the other hand, deep augmentation technique uses large scale pre-trained language models such as BERT (Devlin et al., 2018) and BART (Lewis et al., 2020) to augment the data in a more diverse manner. The deep augmentation technique using BART has been successful in augmenting dataset where the data labels are scarce (Kumar et al., 2020).

2.5 GPT-3 and Few Shot Learning

Large scale language models like GPT-3 and BERT can be easily controlled using natural text making them a good fit for few shot learning. There has been significant development and interest of using prompt-based mechanisms to help large scale pre-trained models to understand the task in hand. (Yoo et al., 2021) proposed a method of first selecting a subset of examples from the dataset, constructing a prompt using the selected set of examples including the meta-information about the dataset and finally extracting augmentation from the language model generation. In our proposed solution, we follow a similar approach of using prompt based NLP to help GPT-3 understand the hash tag generation task in a few shot setting by passing a list of train data points with the target hashtags and a list of testing data points without hashtags. The model is able to generate hashtags for the test data points by understanding the hash tag generation from train data points.

2.6 Hashtag Generation for Images & Videos

As far as we know, extensive work on hashtag generation from images and videos has not been done. HARRISON (Park et al., 2016) is a dataset provided as a baseline to test other models. hashtag generation from images. They consider the hashtag recommendation task as a multi-label classification problem and use CNN architectures to recommend

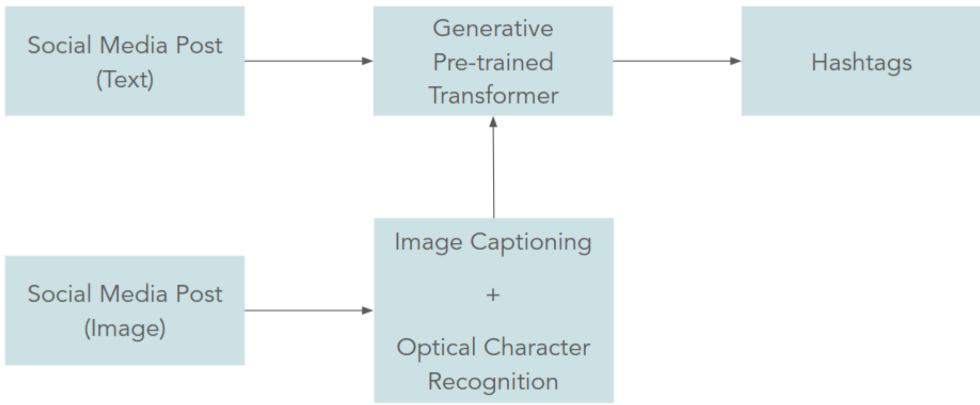

Figure 1: High level diagram for hashtag generation

hashtags.

The Graph Convolution Network based Personalized Hashtag Recommendation (GCN-PHR) model (Wei et al., 2019) was proposed for personalized hashtag recommendations on micro-videos. They use advances in GCN technology to graph complicated user interactions with micro-videos on social media, and generate hashtags for certain parts of the video.

Existing methods use CNN based approaches and may not provide best results. In our work, we will experiment with using image captioning methods first in order to generate captions, on which our text based approach can be extended.

3 Proposed Solution

We propose a framework to generate hashtags from a given social media post using a massive pre-trained language model and augmenting the post for downstream tasks. Figure 1 shows a high-level view of the framework. Our preliminary framework will consist of two main pipelines -

1. Hashtag augmentation for text based posts
2. Hashtag augmentation for image based posts

The hashtag generation module for text will accept a social media post as an input and generate hashtags for that post using a massive pre-trained language model (GPT-3, BART, etc.) in a zero-shot or few-shot setting. Hashtags can be generated for text posts directly using GPT-3. The generated hashtags will be used to augment the social media post in order to provide contextual information and increase outreach. Hashtags can be augmented either as text or as meta-data (in case of images or micro-videos).

For the image pipeline we propose using a combination of image captioning and optical character recognition (OCR) to first generate a caption or text based description of the image, which can then be passed to the PLM (GPT-3) to generate hashtags.

The third pipeline should be one to generate hashtags for micro-videos. Although we haven't implemented the video pipeline in code, we propose to sample a few frames from the video and apply same method as used for images.

Evaluation of our model will combine evaluations of both modules separately. For the hashtag generation module, we will compare against existing hashtag generation and recommendation models using basic metrics such as precision, recall, f1-score, as well as hit-ratio.

Post generation of hashtags, we can annotate social media posts with the hashtags either as text or as meta-data. Addition of hashtags should add contextual data to the post which can improve performance of downstream tasks such as sentiment analysis (Sun et al., 2019), stance detection (Glandt et al., 2021), and toxic comment classification (Caselli et al., 2021). Figure 2 illustrates an example of using hashtag generation and augmentation for a downstream task. We focus our project on using hashtag augmentation to improve post outreach. This can be measured by checking post imprints or views with and without the generated hashtags.

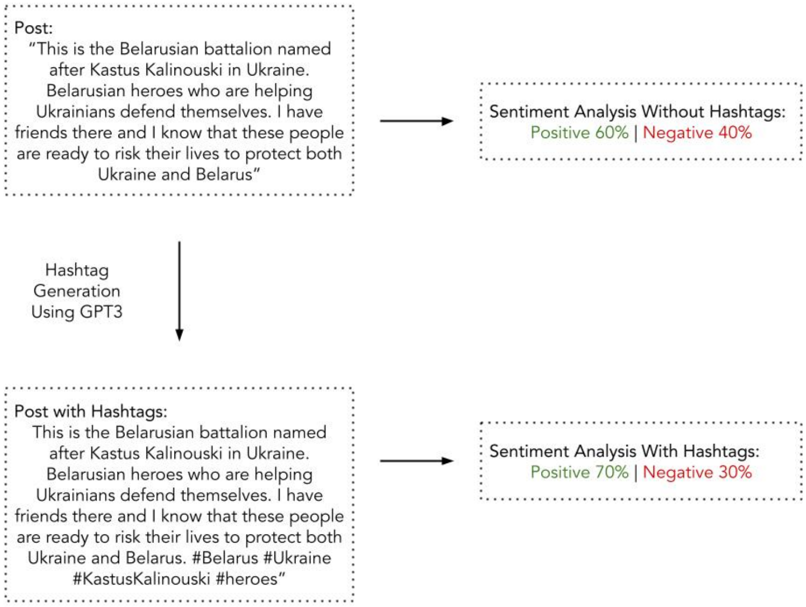

Figure 2: Example of using hashtag generation and augmentation for downstream tasks.

4 Experimental Results

4.1 Datasets

For the midpoint, we compiled a list of 15 popular hashtags on Twitter. Then for each hashtag we found 4 Tweets. For each of those Tweets we extracted the text and all hashtags in that Tweet. Further, we provided our industry mentors with 50 most popular hashtags on Twitter. They are currently creating larger dataset (consisting of only text documents). Currently, our dataset only includes text and hashtags, however, the new dataset will also consist of metadata such as date and location of the user. Additionally, we plan to create a small data set for images and micro-videos possibly from Instagram and/or Tiktok posts.

4.2 Evaluation: The Hit-Ratio Metric

To evaluate the performance of the models we are using the hit-ratio metric. For each tweet in the testing dataset there are a few hashtags. We then generate a few hashtags for each tweet. The length of all hashtags combined is limited to 60 characters. On an average the dataset contains 2.8 hashtags per tweet. On an average GPT3 generates 3.8 hashtags per tweet. If any hashtags generated is in the test set for the tweet it is given a Hit-Rate score (accuracy) of 1, 0 otherwise. i.e.

$$\text{Accuracy} = \begin{cases} 1 \text{ if } [\text{Test Hashtags} \cap \text{Generated Hashtags} \neq \phi] \\ 0, \text{ otherwise} \end{cases}$$

$$\text{Precision} = \frac{|\text{Test Hashtags} \cap \text{Generated Hashtags} \neq \phi|}{|\text{Generated Hashtags}|}$$

$$\text{Recall} = \frac{|\text{Test Hashtags} \cap \text{Generated Hashtags} \neq \phi|}{|\text{Test Hashtags}|}$$

$$\text{F-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.3 Hashtag Generation for Text Documents

For text documents (i.e. tweets containing only text) we performed few-shot learning experiments using various versions of GPT3 to generate hashtags. The results are summarized in table 1.

As we see in table 3, GPT-3 'davinci' and 'curie' were able to generate #cloudsecurity from the Ground Truth Hashtags for the Tweet after few-shot learning, however, other models were not. We see that GPT3 'babbage' and 'ada' were able to produce some meaningful hashtags such as #infosec and #aws, however, we do not reward the models for these hashtags since we want the models to produce exact hashtags from the Ground Truth. It is because even one different character in two hashtags would put them into two different categories on a social media platform.

	Accuracy	*Precision*	*Recall*	*F-Score*
GPT3 ‘davinci’	0.522	0.262	0.282	0.272
GPT3 ‘curie’	0.455	0.235	0.258	0.246
GPT3 ‘babbage’	0.451	0.226	0.240	0.233
GPT3 ‘ada’	0.416	0.212	0.234	0.222

Table 1: Evaluation of GPT-3 Hashtag Generation

	Accuracy	*Precision*	*Recall*	*F-Score*
GPT3 ’davinci’	0.483	0.223	0.128	0.163

Table 2: Evaluation of GPT-3 Hashtag Generation on Image Dataset

4.4 Hashtag Generation for Image Documents

For image documents we first use the ViT + GPT2 image captioning model as well as EasyOCR to convert images to text. We then input the text into GPT3 to generate hashtags. As seen in the table, the model was able to generate relevant hashtags. However, the model performs poorly on some images when the image captioning model performs poorly and there is no text in the post. To evaluate the performance of the model we made a dataset with 149 Twitter posts that contain images. As seen in Table 2, the model achieved an accuracy of 0.483.

Figure 3: Input Image

4.5 Error Analysis

Our experimental results show lower accuracy than expected. Even though generated hashtags look mostly relevant, there were some unrelated outputs and low overlap with ground truth hashtags. Based on our analysis of the outputs, we have observed that the errors are mostly caused due to following reasons -

1. *Variable Context*: Since social media posts taken in isolation have very low context, it becomes hard to distinguish which context is being used for posts with ambiguous references.

2. *Ambiguous Hashtags*: Some hashtags used in ground truth may be ambiguous and the relation to the post may be unclear. Also, several different kinds of hashtags can be used for posts dealing with similar topics. This may affect the accuracy.

3. *Noisy Data*: The dataset taken for experiments was pre-processed but not curated to exclude posts which were click-bait or ads, which affected the pre-trained model.

Table 5 illustrates some examples for each error type we found in the test dataset.

5 Future Work

Currently, we have experimented the hashtag generation module of our framework on a dataset of 2000 tweets pulled from twitter by searching through the list of most popular hashtags (Chris Sabanty, 2019) using an API endpoint provided to us by our industry mentor. We also used a small dataset of 200 tweets with images to evaluate the benefit of passing image captions to improve our module. Our next steps mostly revolve around improving the dataset by adding more filtering and adding more

Tweet	Best practices to secure and manage workloads migrated to Azure - Cloud Adoption Framework — Microsoft Docs
Name	*Hashtags*
Ground Truth	#cloudsecurity #cybersecurity #azure #msftsecurity #cloudfamily #azurefamily
GPT3 'davinci'	#workingforward #technology #software #cybersecurity #machine
GPT3 'curie'	#cloudsecurity #cybersecurity #dataprote #azure
GPT3 'babbage'	#cloudpaa5 #aws #azure35 #dat
GPT3 'ada'	#azure #microsoft #collectiveID #infosec

Table 3: Hashtag Generation for Input Text

Model	*Generated Output*
ViT + GPT2 Image Captioning	an advertisement for a television show
EasyOCR Text	on this dah 14 years ago warriors nation steph currv itroduced hiiself to the iorld hnd scored 40 poiits aghiist hohzhca clutchpoints valdon:
Zero-Shot Hashtag Generation	#Warriors #StephCurry

Table 4: Hashtag Generation for Input Image

contextual information to improve the hashtag generation module.

Broadly, we suggest the following next steps to improve the model and approach based on our error analysis -

1. Before passing the dataset for training to the model, we can filter the dataset by only considering social-media posts which have high engagement in terms of re-post and likes which would filter a large amount of noisy data.

2. We can include more contextual information like videos, recent social-media trends and spatial information like the post creation date which would improve the relevance of generated hashtags for a social media post.

3. We can improve the performance by using better prompt engineering for in-context learning and using the factual knowledge retained in large pre-trained language models like GPT-3 from their training corpora.

4. We can improve our image captioning model by passing more context along with the image which would improve the captions that are generated and than used as context for our hashtag generation module.

5. Experiment with other large PLMs such as Meta AI's OPT (Zhang et al., 2022), or neural topic modellers such as BERTopic (Grootendorst, 2022).

6 Acknowledgements

We would like to thank our industry mentors Keen Sung and Andrew Curran from AuCoDe Inc., and our Ph.D. mentor Subendhu Rongali for helping us with this project.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Error Type	Tweet	Hashtags	Description
Variable Context	Who do you think is the goat of Football?	#Ronaldo #Messi	As we see both example 1 and 2 are discussing about 'football'. However, Tweet 1 is referring to soccer and Tweet 2 is refering to American football. Since two very similar documents produce very different hashtags, it makes it difficult for the model reproduce the correct hashtags with the lack of context.
	Hang on, who do you think the GOAT of football is?	#NFL #Brady #Manning	
Ambiguous Hashtags	Who wants free NFT?	#giveaway #NFT	Users use different hashtags for very similar posts. Due to the ambiguous use of hashtags it becomes difficult for the model to reproduce the correct hashtags.
	Who wants free NFT and crypto?	#Crypto #NFTGiveaway	
Noisy Data	//empty	#gift #affiliate #affiliatemarketing #deal #blogger #business #cryptocurrency #deals #discount #gifts #marketing #shopping #socialmedia #travel #twitter #webtalk #ad #affiliatelink #drinks #alcohol #beer #wine #liquor	The dataset made was very noisy. For example, there were empty tweets with many random hashtags; there were random charcaters with random hashtags; there were hashtags unrelated to the content. Many of these were most likely posted by bots and due to these noisy datapoints, the fine-tuned model sometimes produced unrelated or inaccurate hashtags.

Table 5: Observed errors

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Chris Sabanty. 2019. Popular and most liked hashtags on twitter. [Online; accessed April 27, 2013].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611, Online. Association for Computational Linguistics.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Abhay Kumar, Nishant Jain, Suraj Tripathi, and Chirag Singh. 2019. From fully supervised to zero shot settings for twitter hashtag recommendation.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Qianren Mao, Xi Li, Hao Peng, Bang Liu, Shu Guo, Jianxin Li, Lihong Wang, and Philip S. Yu. 2021. Attend and select: A segment attention based selection mechanism for microblog hashtag generation.

Minseok Park, Hanxiang Li, and Junmo Kim. 2016. Harrison: A benchmark on hashtag recommendation for real-world images in social networks.

Statista. 2022. Share of u.s. population who use social media 2008-2021.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics*, pages 194–206, Cham. Springer International Publishing.

Soheil Tehranipour. 2020. Openai gpt-3: Language models are few-shot learners.

Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. 2011. Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, page 1031–1040, New York, NY, USA. Association for Computing Machinery.

Yue Wang, Jing Li, Irwin King, Michael R. Lyu, and Shuming Shi. 2019. Microblog hashtag generation via encoding conversation contexts.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks.

Yinwei Wei, Zhiyong Cheng, Xuzheng Yu, Zhou Zhao, Lei Zhu, and Liqiang Nie. 2019. Personalized hashtag recommendation for micro-videos.

Yong Wu, Yuan Yao, Feng Xu, Hanghang Tong, and Jian Lu. 2016. Tag2word: Using tags to generate words for content based tag recommendation. In *CIKM 2016 - Proceedings of the 2016 ACM Conference on Information and Knowledge Management*, International Conference on Information and Knowledge Management, Proceedings, pages 2287–2292. Association for Computing Machinery.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyeong Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation.

Qi Zhang, Jiawen Wang, Haoran Huang, Xuanjing Huang, and Yeyun Gong. 2017. Hashtag recommendation for multimodal microblog using co-attention network. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3420–3426.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Xiuwen Zheng, Dheeraj Mekala, Amarnath Gupta, and Jingbo Shang. 2021. News meets microblog: Hashtag annotation via retriever-generator.