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 micro-posts which are constrained to a specific length. These micro-posts cover a wide range of topics ranging from personal, political, social topics to promotional content. Each post can contain a hashtag which consists of a string of characters prefixed with a '#'. Hashtags help reach out to a wider audience by making it easier to search through the content, create instant communities with people of similar interests. Although they are highly effective in helping the user maximize their engagement and fulfill their purpose, only a fraction of the posts contain hashtag. To address this problem we will develop a framework using recent state-of-the-art massive neural text generation models such as BART and GPT3 to generate hashtags given a social-media post, which may contain text, image or videos.

1 Introduction

1.1 Task Description

The main focus of our project will be to automatically generate hashtags for a given social media post using state-of-the-art deep learning models for text generation. A social media post may contain either text, images, or videos. Accordingly, our main contributions to this work would be:

1. Model automated hashtag annotation as a text generation task and use BART to generate hashtags,
2. Experiment with GPT-3 to generate hashtags from given social media post in a zero-shot or few-shot setting,
3. Combine image captioning along with proposed text generation methods to generate hashtags for images,
4. Evaluate performance of these massive PLM text generation models against current SOTA architecture for hashtag generation.

1.2 Motivation and Limitations of Existing Work

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, it becomes imperative to identify techniques that may help maximize a post's reach.

Users may use #hashtags to associate their posts with certain topics and drive engagement. Hashtags can further be used for downstream social media analytics tasks such as search, summarization, sentiment analysis, etc. However, most social-media posts do not have any tagged hashtags (Wang et al., 2011). Hence we propose automatic hashtag generation for social media posts without any associated hashtags.

Previous works on automatic hashtag generation have several limitations such as not being able to generate long phrased hashtags, being highly dependent on large annotated datasets, and not using state-of-the-art techniques for the task at hand. Not much research has been done in the area of hashtag generation from images or micro-videos as well. Our approach attempts to use state-of-the art massive pre-trained language models (PLMs) modified for the task of hashtag generation, as well as adapting these techniques to generate hashtags from images as well.

1.3 Proposed Approach

1. A document may consist of text, images, and hashtags
2. For each document separate the images, text, and hashtags
3. For in-sentence hashtags, remove the '#' character and then treat them as end of the sentence hashtags
4. Train Oscar (Object-Semantics Aligned Pre-training for Vision-Language Tasks)(Li et al., 2020) on the Microsoft COCO dataset(Lin et al., 2014) for image captioning. We plan to use Microsoft COCO dataset as it is rated as the best image captioning dataset by Papers with Code. We intend to use Oscar as our image captioning model because it performs the best on Microsoft COCO dataset with the highest Recall@10 at 98.3
5. Then use Oscar to caption the image and treat the image as a text document from this step onwards
6. Mask the hashtags and use BART to fine tune the transformers to predict the hashtags. Additionally, fine-tune GPT-3 in a zero-shot or few-shot setting to predict hashtags.

Figure 1: Training Model to Generate Hashtags for Text

1.4 Likely Challenges and Mitigations

We are new to a lot of the neural language models like BART, GPT-3 and it might be computationally expensive to fine-tune these large pre-trained models which might become a bottleneck for us. In such cases, we will explore other pre-trained models like BERT for our initial encoding process and a 1D-CNN Classification model for hashtag classification of text . Since, we are planning to build our own-dataset for different social-media platforms, the quality or the scale of the dataset might not be good enough to generate meaningful results. In such cases, we will modify and use pre-existing social media datasets which contain hashtags.

2 Related Work

Our work focuses on automatic hashtag generation using neural text generation. We focus on generating hashtags not only from text-based social media posts, but also from posts containing images or micro-videos.

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 predefined 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 BART (Lewis et al., 2020) and 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.

BART has recently been used for topic modeling as BART-TL (Popa and Rebedea, 2021). BART-TL generates accurate representations of the most important topic terms and candidate labels by fine-tuning on a large number of potential labels generated by state-of-the-art models for topic labeling. A similar approach can be adopted for the task of hashtag generation.

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 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 hashtags.

(Wei et al., 2019) proposes a Graph Convolution Network based Personalized Hashtag Recommendation (GCN-PHR) model 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.

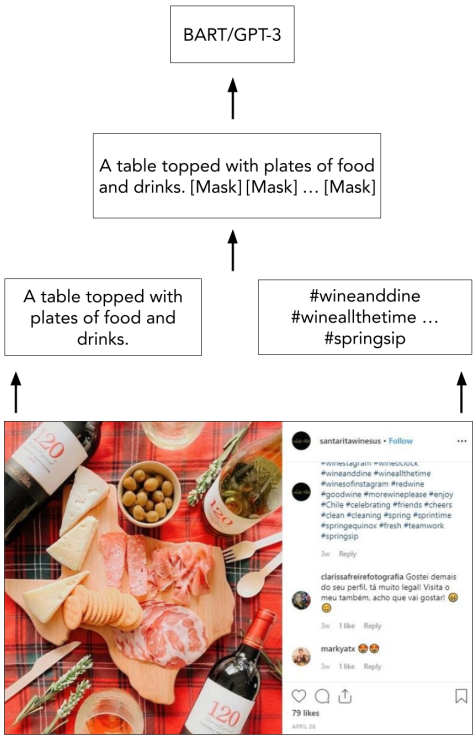

Figure 2: Training Model to Generate Hashtags for Images

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 Experiments

3.1 Datasets

1. *Making our own Dataset with help of Industry Mentors*: Extracting documents (tweets, captions, text, comments, images) from social media platform with hashtags using APIs.
2. *SNAP*: A publicly available dataset by Stanford (Leskovec and Mcauley, 2012) consisting of 476 million tweets by 20 million users with almost 50 million Hashtags. Each data point contains information the tweet's author, time, and content.
3. *Instagram Profiles*: A publicly available dataset consisting of posts collected from

978 Instagram profiles consisting of captions, hashtags, and image URLs (Foco, 2021).

4. *Interactive Facebook Reactions*: A publicly available dataset consisting of posts from several accounts (Woolf, 2016). Each data point consists of the status, link to the images attached in the post, number of likes, comments, and shares.
5. *Microsoft COCO Dataset*: A publicly available large-scale object detection, segmentation, key-point detection, and captioning dataset consisting of 328K images.

3.2 Baselines & Evaluation

Since this in an exploratory research we plan to conduct multiple experiments, and then determine how each model compares to each other. For image captioning, we will explore VisualBERT (Li et al., 2019) as an alternative. To evaluate the performance of our experiments we will use the hit ratio metric (Alsini et al., 2020). For a predetermined k, if in the top k generated hashtags, at least one hashtag match with any ground-truth hashtags, we give it a hit rate of 1, 0 otherwise. The hit-ratio score can be defined as:

$$\frac{\text{sum of hit rate over all tweets}}{\text{number of tweets}}$$

3.3 Updates since the project proposal presentations

We proposed an exploration research about 'Understanding Community Dialects on Twitter'. However, after a discussion with our industry and Ph.D. mentors, we have decided to head in a different direction to work on 'Generating Hashtags for Different Social Media Platforms'.

3.4 Software

Our work will involve two major parts – (a) dataset building and, (b) model development and experimentation. Dataset building will involve coding using various social media APIs such as the Twitter API, Instagram Basic Display API, and Facebook Basic Display API. During model development, we will use Pytorch along with the 🤗 Transformers. HuggingFace Transformers has an implementation and pretrained version of BART. We will also use GPT-3 API for zero-shot and few-shot approaches. We will maintain our code on GitHub and run models using UMass Amherst's Unity cluster.

3.5 Timeline

1. *2/11 - 2/18:* Improve and finalize project proposal based on feedback received by Industry mentors, Ph.D. mentor, and instructors
2. *2/19 - 2/26:* Collect and build all detests necessary. Explore all pre-existing datasets
3. *2/27 - 3/20:* Implement Baselines and conduct experiments
4. *3/21 - 3/25*: Work on midpoint presentation and submission

3.6 Challenges

We want to extract documents from each social media platforms keeping them as consistent as possible across the platforms in terms on temporal features and distribution of the topics. We believe this is something which can be challenging to achieve.

3.7 Contingency Plan

In case extracting data from social media platforms ends up causing insurmountable issues, we plan to use pre-existing datasets. Backup experiment – treating the task as a classification problem where images/text are documents to be classified and hashtags are the labels. Additionally, if processing images and micro-videos proves to be too resource and time intensive, we plan to focus solely on text data.

4 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 proposal.

References

Areej Alsini, Du Q. Huynh, and Amitava Datta. 2020. Hit ratio: An evaluation metric for hashtag recommendation.

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.

Pio Foco. 2021. Instagram profiles - dataset by prsr.

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

Jure Leskovec and Julian Mcauley. 2012. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

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.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft coco: Common objects in context.

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.

Cristian Popa and Traian Rebedea. 2021. BART-TL: Weakly-supervised topic label generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1418–1425, Online. Association for Computational Linguistics.

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

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.

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

Max Woolf. 2016. Interactive facebook reactions.

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.

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.

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