Hashtag-Guided Low-Resource Tweet Classification

Shizhe Diao*
The Hong Kong University of Science and Technology
sdiaoaa@connect.ust.hk

Sedrick Scott Keh*
Carnegie Mellon University
skeh@cs.cmu.edu

Liangming Pan
University of California, Santa Barbara
liangmingpan@ucsb.edu

Zhiliang Tian
The Hong Kong University of Science and Technology
tianzhilianghit@gmail.com

Yan Song
University of Science and Technology of China
clksong@gmail.com

Tong Zhang
The Hong Kong University of Science and Technology
tongzhang@ust.hk

ABSTRACT

Social media classification tasks (*e.g.*, tweet sentiment analysis, tweet stance detection) are challenging because social media posts are typically short, informal, and ambiguous. Thus, training on tweets is challenging and demands large-scale human-annotated labels, which are time-consuming and costly to obtain. In this paper, we find that providing hashtags to social media tweets can help alleviate this issue because hashtags can enrich short and ambiguous tweets in terms of various information, such as topic, sentiment, and stance. This motivates us to propose a novel **Hash**tag-guided **T**weet Classifi**cation** model (**HashTation**), which automatically generates meaningful hashtags for the input tweet to provide useful auxiliary signals for tweet classification. To generate high-quality and insightful hashtags, our hashtag generation model retrieves and encodes the post-level and entity-level information across the whole corpus. Experiments show that HashTation achieves significant improvements on seven low-resource tweet classification tasks, in which only a limited amount of training data is provided, showing that automatically enriching tweets with model-generated hashtags could significantly reduce the demand for large-scale human-labeled data. Further analysis demonstrates that HashTation is able to generate high-quality hashtags that are consistent with the tweets and their labels. The code is available at https://github.com/shizhediao/HashTation.

CCS CONCEPTS

• **Information systems** → **Web mining**.

KEYWORDS

social media analysis, tweet classification, hashtag generation, low-resource classification

*Both authors contributed equally to this research.

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583194

ACM Reference Format:
Shizhe Diao, Sedrick Scott Keh, Liangming Pan, Zhiliang Tian, Yan Song, and Tong Zhang. 2023. Hashtag-Guided Low-Resource Tweet Classification. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3543507.3583194

1 INTRODUCTION

Table 1: Examples of how hashtags can provide auxiliary information for better tweet classification.

Original Input	Generated Hashtags	New Hashtag-Guided Input
Abortion IS NOT a political issue. It is a MORAL issue	AllLivesMatter , ProLife	Abortion IS NOT a political issue. It is a MORAL issue #AllLivesMatter #ProLife
Twitter making everybody mad. It's hilarious	Twitter, hilarious	#Twitter making everybody mad. It's #hilarious
He's the GOAT for sure!	GOAT, NBAFinals	He's the #GOAT for sure! #NBAFinals

Tweet Classification (TC) is an essential task in social media content analysis, which aims to analyze user behaviors and attitudes on Twitter. Typical tasks in this area include stance detection [36], sentiment analysis [37], and hate speech detection [4]. However, these tasks are often difficult because tweets are usually informal, idiosyncratic, and short in length and thus provide limited and ambiguous information. Additional context and background knowledge are often needed to understand the content of a tweet better. Due to this lack of information and the ambiguous nature of tweets, we often need to label a large-scale training corpus in order to train a satisfactory TC model [3, 37]. However, the rapidly changing and evolving nature of social media content makes it challenging to annotate in-domain training data in a timely manner. Furthermore, data annotation is time-consuming and costly. To address the above challenges, we propose a model, HashTation, with two novel features: 1) it can automatically enrich the content of social media tweets by *hashtag generation*, and 2) the hashtag-enriched tweet classification model works well under the *low-resource* setting in which only a limited amount of labeled data is available.

Hashtags are commonly contained within tweets or appended to the end of tweets. They not only facilitate rapid lookup for specific themes or web contents, but also contain important information that helps to enrich and disambiguate the contents of tweets. As exemplified in Table 1, we can hardly understand the topic and sentiment of the tweet "Abortion IS NOT a political issue. It is a MORAL issue." without its hashtag "#RoeVWade". Our pilot study (Table 2)