INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS

Name of the Student	:	Durgesh Kumar
Roll Number	:	126101002
Programme of Study	:	Ph.D.

Thesis Title: Significance of Hashtags for Improved Topic Modeling on Tweets

Name of Thesis Supervisor(s)	:	Dr. Sanasam Ranbir Singh
Thesis Submitted to the Department/ Center	:	Computer Sicence and Engineering
Date of completion of Thesis Viva-Voce Exam	:	16/03/2022
Key words for description of Thesis Work	:	Topic Modeling, Hashtags, tweet expansion, hashtag prioritization, Social Network Anlaysis

SHORT ABSTRACT

With the increase in Twitter's popularity, topic modeling on Twitter has become an important problem with applications in diverse fields such as text summarization, document clustering, information retrieval, and sentiment analysis. The short and noisy tweets with informal writing style make topic modeling on tweets more challenging due to increased data sparsity and under-specificity. Latent Dirichlet Allocation (LDA), one of the widely used topic models, suffers from data sparsity and under-specificity. Researchers have tried to counter the data sparsity and under-specificity in tweets by adding related content from external sources such as News pages and Wikipedia or pooling related tweets to pseudo documents. Adding the content from external resources is non-trivial due to differences in writing styles and vocabulary. Moreover, Topic modeling on pooled documents may lose the distribution of topics over the individual tweet and increase the corpus size due to duplicate tweets in different pools. From earlier studies and our preliminary investigation, hashtags are found to provide necessary meta-information in linking tweets to the underlying topics. Motivated by the above observation, this thesis proposes two approaches to counter the data sparsity and under-specificity in tweets for topic modeling tasks: i) expanding tweets with semantically related hashtags, and ii) prioritization of selected hashtags. From various experimental results, it is evident that our proposed methods enhance the topic modeling performance either by i) tweet expansion with semantically related hashtags or ii) incorporating prioritized hashtags in LDA. Furthermore, this thesis investigates the effect of LDA in relation prediction as a case study by exploiting topic and entity relation. It is observed that event-centric relations are effectively predicted using topic modeling over news articles.