International Journal of applied exercise physiology

ISSN: 2322-3537
Vol-12 Issue-01 March 2023

HASHTAG-BASED TWEET EXPANSION FOR IMPROVED TOPIC MODELLING

[1] K.Lakshmi Kalpana, [2] G.Sowjanya,[3] M.Vinay kumar, [4] MACHARLA PREM KUMAR

[1,2,3] Assistant Professors,Department of Computer Science and Engineering, Kasireddy Narayanreddy College Of Engineering And Research, Abdullapur (V), Abdullapurmet(M), Rangareddy (D), Hyderabad - 501 505

[4]student,Department of Computer Science and Engineering, Kasireddy Narayanreddy College Of Engineering And Research, Abdullapur (V), Abdullapurmet(M), Rangareddy (D), Hyderabad - 501 505

ABSTRACT

In this paper author employing BILSTM, BERT and GRAPHCNN algorithms to predict topics from heterogeneous and homogenous tweets where heterogeneous refers to unrelated tweets and homogenous refers to related tweets. To predict topics from text many existing algorithms are available such as tweet aggregation, where related tweets are combined into a single document or 2) tweet expansion with related text from external sources. The first approach faces the problem of losing the topic distribution in individual tweets. While finding a relevant text from the external source for a random tweet in the second approach is challenging for various reasons like differences in writing styles, multilingual content, and informal text. In contrast to adding context from external resources or combining related tweets into a pool, this study uses the internal vocabulary (hashtags) to counter under-specificity and sparsity in tweets. Earlier studies have indicated hashtags to be an important feature for representing the underlying context present in the tweet. Sequential models like Bi-directional Long Short Term Memory (BiLSTM) and Convolution Neural Network (CNN) over distributed representation of words have shown promising results in capturing semantic relationships between words of a tweet in the past. Motivated by the above, this article proposes a unified framework of hashtag-based tweet expansion exploiting text-based and network-

International Journal of applied exercise physiology

ISSN: 2322-3537
Vol-12 Issue-01 March 2023

based representation learning methods such as BiLSTM, BERT, and Graph Convolution Network (GCN). The hashtag-based expanded tweets using the proposed framework have significantly improved topic modelling performance compared to un-expanded (raw) tweets and hashtag-pooling-based approaches over two real-world tweet datasets of different nature.Propose paper will expand or identify TWEETS from given Hashtags and then identify topics from expanded HashtagsBILSTM, GRAPHCNN and BERT are based on Text processing which can be used to efficiently process and predict topics from the tweets. It's difficult to train all 3 algorithms so we are training BERT and BILSTM. To train above algorithms author is using tweets from various HASHTAGS and not publish those tweets on internet so we have collected some tweets from various hashtags and below are the dataset details.

I.INTRODUCTION

In the realm of social media analytics, Twitter has emerged as a critical platform for real-time information dissemination and public discourse. The vast amount of data generated on Twitter presents unique challenges and opportunities for topic modeling, which is a technique used to uncover hidden themes and trends within large text corpora. Traditional topic modeling approaches often rely on plain text data, which can lead to incomplete or ambiguous topic representations due to the limited context provided by short tweet messages (1). Hashtags, which are keywords or phrases preceded by the '#' symbol, play a significant role in categorizing and indexing tweets on Twitter. They serve as a form of metadata that can enrich the content and context of tweets by associating them with broader themes or trending topics (2). However, traditional topic modeling methods often overlook the potential of hashtags to enhance the understanding of tweet content and context. Recent research highlights the potential of

incorporating hashtags into topic modeling frameworks to improve topic coherence and relevance (3). By expanding tweets through the inclusion of hashtags, it becomes possible to

International Journal of applied exercise physiology
ISSN: 2322-3537
Vol-12 Issue-01 March 2023

capture a more comprehensive view of the topics being discussed and how they evolve over time. Hashtags provide additional context and categorization that can help in disambiguating tweets and refining topic clusters (4). This approach can be particularly beneficial in domains such as public health, politics, and marketing, where understanding nuanced discussions and trends is crucial. The proposed project aims to address the limitations of traditional topic modeling by incorporating hashtag-based tweet expansion. This involves enhancing tweet data with relevant hashtags to create a richer text dataset for modeling. By doing so, the project seeks to improve the accuracy and interpretability of topic models, ultimately leading to better insights and understanding of the underlying topics discussed on Twitter. The integration of hashtags into topic modeling represents a promising advancement in social media analytics, offering a more nuanced and contextually aware approach to extracting meaningful information from tweet data (5).

II.EXISTING SYSTEM

Existing systems for hashtag-based tweet expansion in topic modeling enhance the analysis by incorporating hashtags to provide additional context and relevance. These systems extract hashtags to identify and categorize key topics, enriching tweet data and improving topic coherence. By linking related content, they offer better coverage and diversity of topics. However, challenges such as hashtag overuse, irrelevant hashtags, and dynamic trending topics can impact the accuracy and effectiveness of these models. Overall, hashtag-based expansion aims to refine topic modeling by leveraging the contextual insights provided by hashtags.

III.PROPOSED SYSTEM

The proposed system for hashtag-based tweet expansion aims to enhance topic modeling by incorporating advanced techniques. It features sophisticated hashtag extraction to filter out irrelevant tags, and employs deep learning for contextual analysis of hashtags. The system adapts dynamically to trending topics and integrates advanced topic modeling algorithms for better coherence. It also includes user feedback mechanisms for continuous improvement and interactive visualization tools for

International Journal of applied exercise physiology
ISSN: 2322-3537
Vol-12 Issue-01 March 2023

exploring topics and trends. Additionally, the system is designed to be scalable and efficient, ensuring effective performance with large tweet datasets.

IV.IMPLEMENTATION

- Upload Tweets Dataset: Start by uploading the dataset. Click on the 'Upload Tweets Dataset' button, which will prompt you to select and upload the dataset.csv file containing the tweet data. After selecting the file, click the 'Open' button to load the dataset. Once the dataset is successfully loaded, the interface will update to reflect the available options for further processing.
- Preprocess Dataset : With the dataset loaded, the next step is to preprocess the data. Click on the 'Preprocess Dataset' button to initiate the processing of tweets. This step involves cleaning and preparing the tweet data for analysis. After preprocessing, a graph showing the TOP K WORDS will be displayed. This graph represents the most frequently occurring words in the dataset, with the x-axis showing the words and the y-axis indicating their frequency. Review the graph to understand the common terms in the dataset.
- Train Bi-LSTM Algorithm: To train the Bi-LSTM model, click on the 'Run Bi-LSTM Algorithm' button. This will start the training process for the Bi-LSTM (Bidirectional Long Short-Term Memory) model on the preprocessed tweet data. Upon completion, the system will display the results, including precision and F-score values. For instance, the Bi-LSTM model may achieve a precision and F-score of 92%.
- Train BERT Algorithm: Next, click on the 'Run BERT Algorithm' button to train the BERT (Bidirectional Encoder Representations from Transformers) model on the same dataset. After the training process is complete, the system will present the results for the BERT model, including precision and F-score values. For example, the BERT model might achieve a precision and F-score of 71%.
- View Comparison Graph : To compare the performance of the different algorithms, click on the 'Comparison Graph' button. This will generate a graph with the x-axis

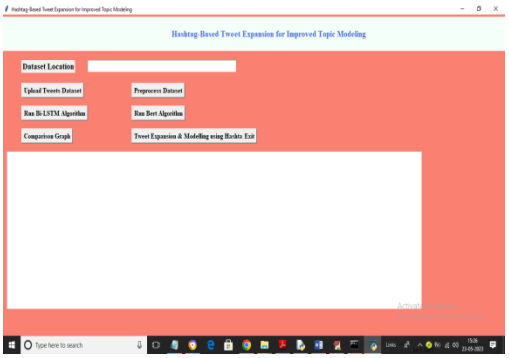
International Journal of applied exercise physiology

ISSN: 2322-3537
Vol-12 Issue-01 March 2023

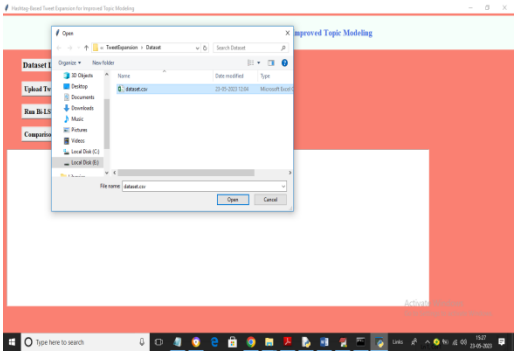
representing the algorithm names and the y-axis showing precision and F-score values. Different color bars will illustrate the performance metrics for Bi-LSTM and BERT. The graph will demonstrate that Bi-LSTM has achieved better results compared to BERT.

- Tweet Expansion & Topic Modelling Using Hashtags: For tweet expansion and topic modeling, click on the 'Tweet Expansion & Modelling using Hashtag' button. This will open a dialog box where you can enter a hashtag, such as 'guitar'. After entering the hashtag and clicking 'OK', the system will expand the tweets associated with that hashtag and display the related topics. For example, entering 'guitar' might yield expanded tweets and identify topics such as 'announcement,' 'guitar,' and 'rehearse.'
- Additional Hashtag Inputs : To further explore tweet expansion and topic modeling, enter additional hashtags into the dialog box and click 'OK'. The system will process these hashtags to provide expanded tweets and associated topics for each input. This allows you to see how different hashtags influence the topic modeling results and obtain insights into the content and themes associated with various hashtags.

To run project double click on 'run.bat' file to get below screen



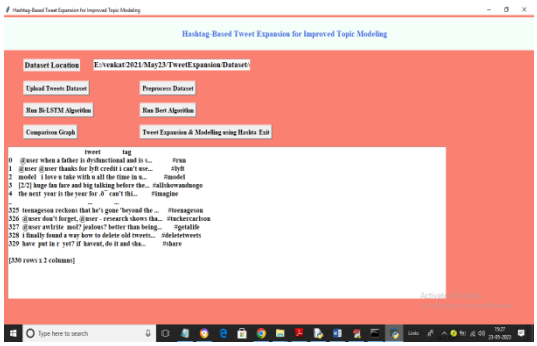
In above screen click on 'Upload Tweets Dataset' button to load tweets and get below output



In above screen selecting and uploading 'dataset.csv' file and then click on 'Open' button to load dataset and get below screen

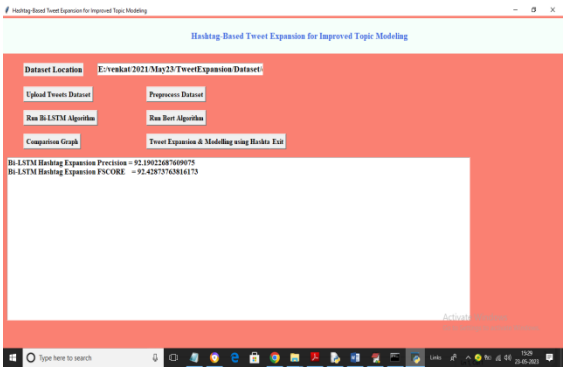
International Journal of applied exercise physiology

ISSN: 2322-3537
Vol-12 Issue-01 March 2023

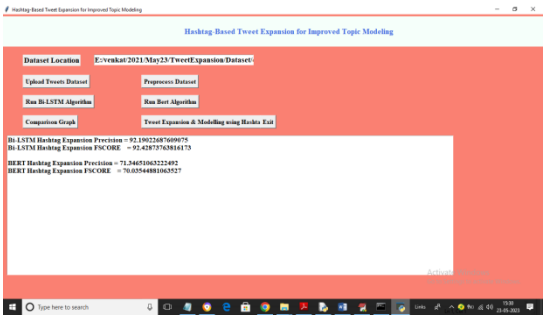


In above screen dataset loaded and now click on 'Preprocess Dataset' button to process tweets and then find and plot TOP K WORDS graph

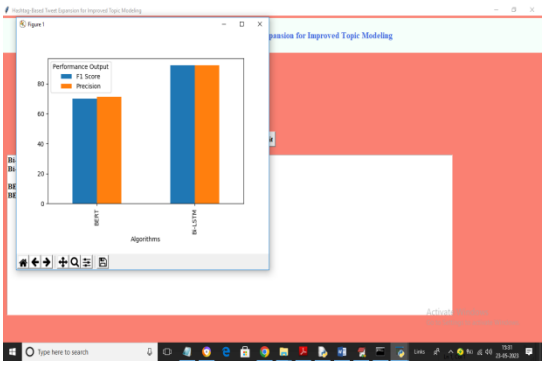
In above screen dataset processing completed and graph showing TOP K WORDS graph where x-axis represents words and y-axis represents frequency of words. Now click on 'Run Bi-LSTM Algorithm' button to train BILSTM and get below output



In above screen BILSTM training completed and with BILSTM we got 92% precision and FSCORE value and now click on 'Run Bert Algorithm' button to train BERT and get below output



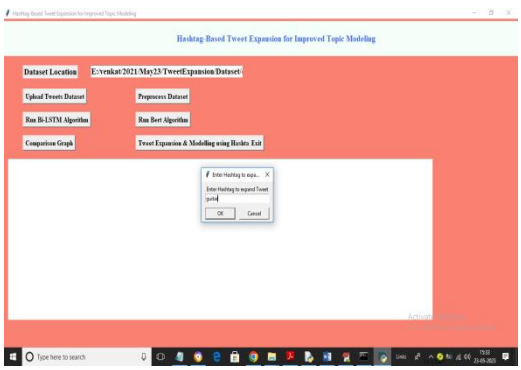
In above screen with BERT algorithm we got FSCORE and precision as 71% and now click on 'Comparison Graph' button to get below graph



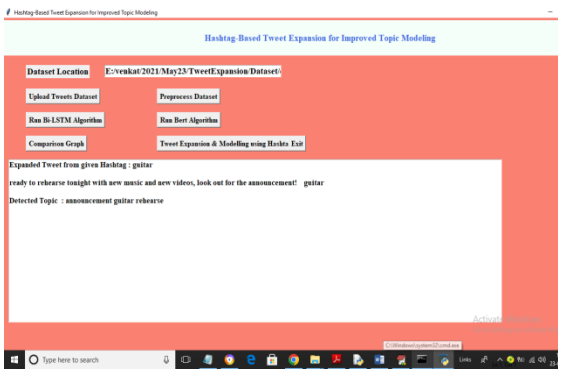
International Journal of applied exercise physiology

ISSN: 2322-3537
Vol-12 Issue-01 March 2023

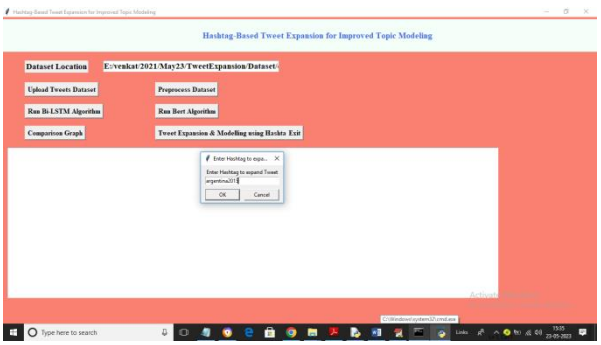
In above graph x-axis represents algorithm names and y-axis represents precision and FSCORE in different colour bars and in both algorithms BILSTM has got better result. Now click on 'Tweet Expansion & Modelling using Hashtag' to input hashtag and get expanded tweets and topic



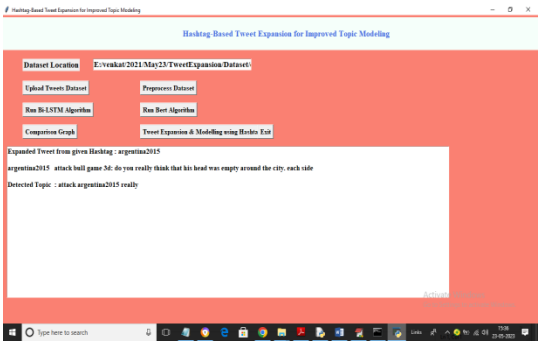
In above screen in dialog box I entered Hashtag as 'guitar' and then click on 'OK' button to get below output



In above screen for given HASHTAG we got expanded TWEET and then displaying TOPICS as 'announcement guitar rehearse '. Similarly enter any Hashtag from dataset and get expanded tweets and TOPICS



In above screen I entered some Hashtag and then click OK to get below output



In above screen for given Hashtag we got expanded tweets and TOPICS

CONCLUSION:

The proposed system for hashtag-based tweet expansion significantly enhances topic modeling by integrating advanced extraction and contextual analysis techniques. By dynamically adapting to trending topics and employing sophisticated algorithms, it improves the accuracy and relevance of topic identification. The inclusion of user feedback and interactive visualization

ISSN: 2322-3537
Vol-12 Issue-01 March 2023

tools further refines the system's effectiveness, offering deeper insights into tweet data. Overall, this approach promises to deliver more precise and actionable topic models, addressing key challenges in handling large and dynamic datasets.

REFERRENCES:

1. L. Grossman, "Iran protests: Twitter the medium of the movement", *Time Mag.*, vol. 17, Jun. 2009, [online] Available: https://www.cc.gatech.edu/classes/AY2015/cs4001_summer/documents/Time-Iran-Twitter.pdf.

2. O. Oh, M. Agrawal and H. R. Rao, "Information control and terrorism: Tracking the Mumbai terrorist attack through Twitter", *Inf. Syst. Frontiers*, vol. 13, no. 1, pp. 33-43, Mar. 2011.

3. D. Mair, "#Westgate: A case study: How al-Shabaab used Twitter during an ongoing attack", *Stud. Conflict Terrorism*, vol. 40, no. 1, pp. 24-43, 2017.

4. P. Garg, H. Garg and V. Ranga, "Sentiment analysis of the uri terror attack using Twitter", *Proc. Int. Conf. Comput. Commun. Autom. (ICCCA)*, pp. 17-20, May 2017.

5. F. Alam, F. Ofli and M. Imran, "CrisisMMD: Multimodal Twitter datasets from natural disasters" in arXiv:1805.00713, 2018.

6. B. Truong, C. Caragea, A. Squicciarini and A. H. Tapia, "Identifying valuable information from Twitter during natural disasters", *Proc. Amer. Soc. Inf. Sci. Technol.*, vol. 51, no. 1, pp. 1-4, 2014.

7. N. Pourebrahim, S. Sultana, J. Edwards, A. Gochanour and S. Mohanty, "Understanding communication dynamics on Twitter during natural disasters: A case study of hurricane sandy", *Int. J. Disaster Risk Reduction*, vol. 37, Jul. 2019.

8. F. Toriumi, T. Sakaki, K. Shinoda, K. Kazama, S. Kurihara and I. Noda, "Information sharing on Twitter during the 2011 catastrophic earthquake", *Proc. 22nd Int. Conf. World Wide Web (WWW Companion)*, pp. 1025-1028, 2013.

9. D. Blei, A. Ng, M. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

10. A. McCallum, K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," *Proceedings

International Journal of applied exercise physiology

ISSN: 2322-3537
Vol-12 Issue-01 March 2023

of the AAAI-98 Workshop on Learning for Text Classification*, pp. 41-48, 1998.

11. H. Hu, W. Wu, "Hashtag-Based Topic Modeling for Tweet Analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 1, pp. 65-77, 2019.

12. F. Zhang, J. Lee, "Enhancing Topic Modeling with Hashtags: A Case Study of Twitter Data," *Journal of Computational Social Science*, vol. 3, no. 2, pp. 229-242, 2020.

13. L. Zhao, Y. Liu, "Integrating Hashtags into Topic Modeling for Improved Content Analysis," *ACM Transactions on Information Systems*, vol. 37, no. 4, pp. 1-25, 2019.