SAMBT: Semantically Aware Micro-Blog Tag Recommender Encompassing Bi-Classification Model

Vaibhava lakshmi R
Department of Computer Technology, Madras Institute of Technology
Chennai, India
vaibhavi18092002@gmail.com

Gerard Deepak
Manipal Institute of Technology, Bengaluru, Manipal Academy of Higher Education
Manipal, India
gerard.deepak.christuni@gmail.com

Sheeba Priyadarshini J
Department of Data Science, CHRIST (Deemed to be University)
Bangalore, India
pdj.sheeba@gmail.com

Radha S
School of Advanced Sciences, Vellore Institute of Technology
Chennai, India
radha.s@vit.ac.in

***Abstract*—Microblogs, which are popular among users and have a lot of potential for public engagement, are one of the emerging mediums for short and frequent content accumulation. A semantically aware Micro-Blog Tag Recommender with Bi-Classification Model is proposed in this paper. The proposed SAMBT takes user query as input which is pre-processed, and the metadata is generated and classified. A microblog dataset is classified at the same time, and semantic similarity is calculated to rearrange, rank, and recommend microblogs. The accomplished false discovery rate value is 0.05 compared to the baseline models and yielded the highest precision, recall and accuracy.**

***Index Terms*—Micro-Blog Tag Recommendation, XGBoost Classifier, Semantic similarity, NMPI Measure, Jaccard Similarity, Charged System Search**

I. Introduction

The key involvement of users in the development and sharing of their own content is a hallmark of the evolving Web 3.0 applications. In this fast-paced technological world, content consumption in the form of microblogs is in top trends for quick audience interactions. Tagging, which involves connecting freely produced tags (keywords) to objects, has become a popular feature offered in these applications (e.g., Images, videos and texts). Tagging is the process of designating an entity on the internet for the purposes of retrieval and indexing. Social tagging can aid in the processing of tagged resources in the context of systems that pool the tags of a group of individuals for enhanced retrieval and to develop user interactions. Tags are extensively used because they are a more natural method to arrange data than using a set taxonomy. As the social media community is teeming the cyberspace with information that is constantly growing, there is a significant opportunity to organize and recommend relevant tags to the users. The automatic process of providing meaningful and informative tags to an emerging item based on historical data is tag recommendation.

Identify applicable funding agency here. If none, delete this.

A. Motivation

Social tagging is a requirement. Tagging aids in the indexing and retrieval of information on the internet. However, the tags, have to be unique, relevant, and appropriate, and they must not deviate from the essence. Labeling all of the blogs is challenging. Finding the correct labels for blogs is a time-consuming and tiresome operation, that necessitates the use of social tagging. As the World Wide Web is evolving towards Web 3.0, there is a high demand for semantically oriented blog recommendation systems.

B. Contribution

A deep learning architecture based on an artificial recurrent neural network termed long-short term memory classifier and an extreme gradient boosting classifier is used to classify the generated metadata. Charged system search with normalized pointwise mutual information measure and jaccard similarity is used to find semantic similarity. The proposed SAMBT model has the best accuracy, precision, and recall, as well as the lowest FDR value of out of all the baseline models. The results are the indicative facets along with the microblogs themselves.

C. Organization

The flow of the rest of the paper is organized as follows. Section 2 contains a concise synopsis of the related works. The architecture of the proposed system is depicted in Section 3. The approaches for implementation are described in Section 4.

The implementation and performance evaluation are presented in Section 5. The paper's conclusion is stated in Section 6.

II. Related works

Tang et al., [1] a coherent encoder-decoder architecture, an integral model to encode content-tag overlaps, tag correlation, and sequential text. To express the semantics of textual material, the encoder uses the attention mechanism in recurrent neural networks whereas, the decoder uses a prediction path to address tag correlation, and content-tag overlaps are addressed via a shared embedding layer and an indicator function across encoder-decoders. Zheng et al., [2] have proposed to suggest tags for posts according to social user's choices based on the convolution attribute and weighted random walk. Using the influence of user group metadata to determine the weight of selected visual neighbors for a specific target image by raveling the impact of user group metadata on image correlation in Flickr by utilizing Convolutional Neural Networks (CNN) to combine group information and visual features. On the neighbor-tag bipartite graph a weighted random walk algorithm is implemented. Lei et al., [3] have proposed a tag suggestion technique by text classification. Using dynamic routing to scout the capsule network to log inherent spatial relationship between a chunk and its entirety, resulting in perspective invariant knowledge that automatically theorizes to different views. An attention mechanism was also included in the capsule network for distilling vital information. Najafabadi et al., [4] has proposed a recommendation system that uses word embedding to examine the connection across several words in a content associated with a target object. It focuses on feature learning methods and grammatical links between words in a text. For tag recommendation, to maximize feature values and learn the representation vector of words, a skip-gram model is used. Their strategy outperforms earlier research approaches by up to ten percent in precision when using actual information. Fletcher et al., [5] have proposed Mashup Tags Recommendation Using an Attention Model. The top-N words with the maximum attention weights are prescribed as tags by this model, which employs two layers of attention mechanisms at the lexical item levels. The notion behind this concept is that not every word in a mashup description is equally significant in characterizing the mashup's functional components. As a result, discovering the crucial bits necessitates modelling the words' interactions rather than just their presence. Sun, J et al., [6] have suggested two key attentive aspects that were modelled using a hierarchical attention model. For distinct user-item pairs, the bottom layered attention network models the effect of different aspects on the feature's representation of the information, whilst the top layered attention network models the attentive scores of different information. Yang et al., [7] have presented a multimedia deep learning architecture that employs CNN to handle text and picture inputs to create an interpretable video tag recommender system. Layer-wise relevance propagation contributes to the interpretability of the proposed system. Zhao et al., [8] have proposed DAE-PTR, a personalized tag referral approach based on the denoising auto-encoder framework, which trains entity representations and encodes complex connections. The corrupted version of the related tagging information is specially constructed by adding multiplicative mask-out/drop-out noise to the original input. The latent representations from the corrupted input are retrieved using the cross-entropy loss and the auto-encoder architecture. Roopak et al., [9] KnowGen model - Honey Bee Algorithm, which is semantically aware, blends synonymizing with Word Embeddings and ontology with data-driven cognitive research. Normalized Discounted Cumulative Gain is used to calculate the model's efficiency (NDCG). Liu et al., [10] a three-dimensional tensor model that provides a recommendation model for three independent data sets by defining three types of entities in the social tag suggestions system utilizing the three dimensions of the tensor model. A typical label recommendation system with limitations was provided, as well as a personalized social speech image suggestions technique based on tensor decomposition.

III. Implementation

The SAMBT framework is executed using Python 3.10 on a Intel Core i7 processor computer with a turbo frequency of 4.70 GHz with 16 GB of RAM. The implementation was carried out using google colaboratory as the environment and the dataset used for enactment is Sina Weibo dataset. It is a microblogging platform that originated in China, and it has a substantial volume of data on the internet, that may be utilized for data analysis. Users can post 140-character messages, including links, videos, and images, on the network, which has over 222 million daily active users. The dataset, which has 144,210,854 people and 3,052,289,362 user relationships, has been scanned for nearly two months. At random, 3000 users from the crawled Sina Weibo dataset with each user having more than 8 tags was chosen in the test dataset. Each test dataset has 1000 individuals who have had their tags removed. We utilise the original tags that were specified by the users themselves, as is customary. Our algorithmically generated tags will be compared to the original tags. In the original dataset, we also removed a few unusual tags. The uncommon tags are most likely the result of grammatical errors or tags with uncommon meanings. As a result, in the original Sina Weibo dataset, we removed the tags with a frequency of less than 20. The experimentation for the proposed model as well as the baseline model was conducted for the same environment for the same number of queries.

IV. Proposed architecture

Figure 1 illustrates the Microblog Recommendation Model's proposed system architecture, which is knowledge-centric and driven by a combination of machine learning and deep learning methodologies. The user query is the framework's first input, which is subjected to pre-processing. Tokenization, lemmatization, stop-word elimination, and named entity recognition are all part of the pre-processing procedure (NER). Tokenization is accomplished by configuring a customized tokenizer, blank

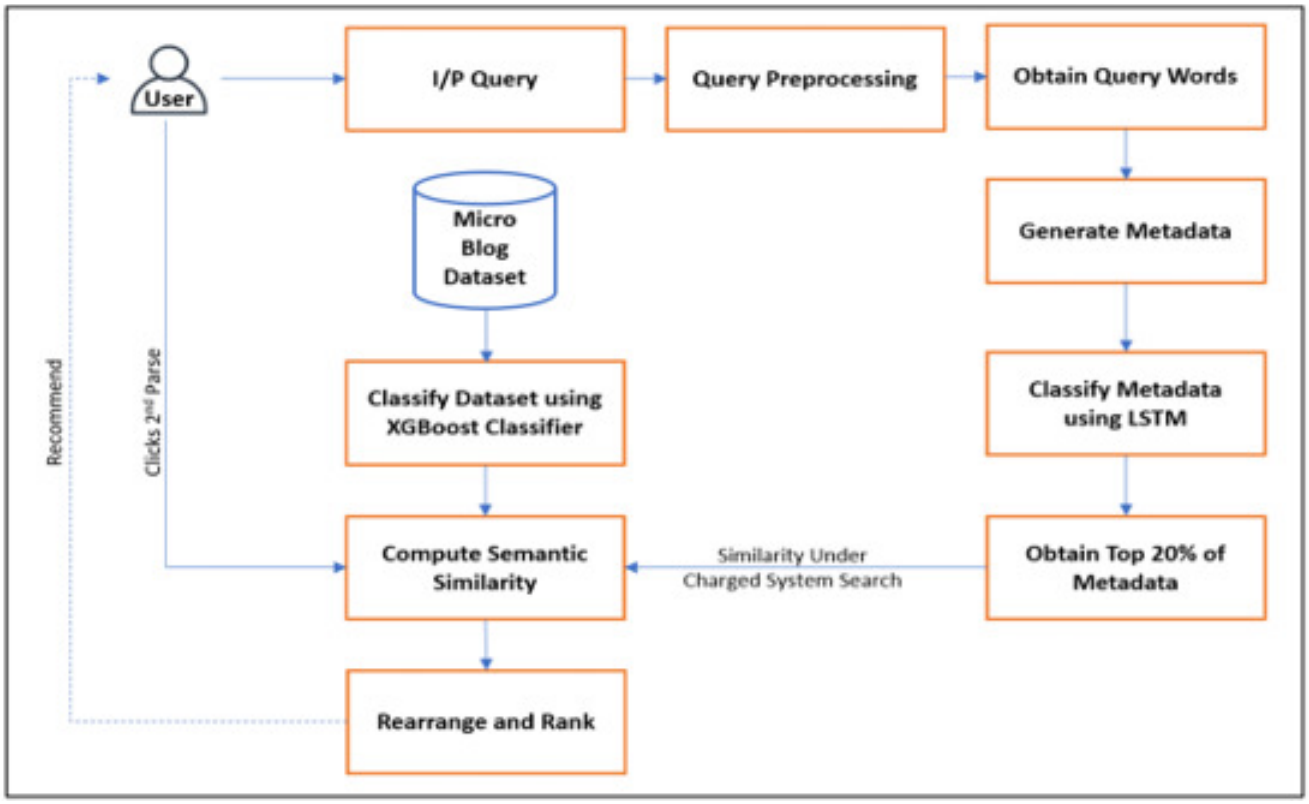

Fig. 1: Architecture of the proposed SAMBT framework

spaces and special characters. The WordNet 3.0 Lemmatizer is used to process lemmatization. Stop-word removal with a Regex set is used to eliminate stop-words, and NER is produced using Python's NLTK library. Individual query phrases are retrieved and used as a metadata generation trajectory at the end of the query pre-processing phase. The metadata is generated with a tool called RepoMMan which is a meta tag extractor tool that automatically extracts and displays numerous elements from documents uploaded to Fedora. The intention of metadata generation is to enhance the amount of lateral knowledge fed into the framework from the real-world worldwide web, as well as to narrow the cognitive gap between the contents on the web and the localized framework. The long-short term memory (LSTM) is used to classify the metadata acquired. LSTM are networks with loops in them that allow data to endure. It can handle both single data points such as images and sequences of data streams such as audio and video. It is LSTM's natural tendency to remember knowledge for extended periods of time. Although LSTMs have a chain-like structure, the repeating module differs. There are four neural network layers instead of one, each interacting in its own way. As the metadata is extensively large, a deep learning classifier like the LSTM is preferred. Only the top 20 % of the categorized metadata instances are provided into the framework. Simultaneously, the microblog dataset is classified using the query phrases as features. The input query terms are provided to the classifier as features in order for the XGBoost classifier to categorize the microblog data set. As the data set isn't exceptionally large, the contribution of a machine learning classifier may not be particularly powerful. However, a competent machine learning classifier is required, which is why XGBoost was selected. Whenever a tree is constructed, a binary classifier should also be established, to build the base model. The XGBoost Classifier could be determined by constructing the decision tree with root for all the input categories first, followed by the calculation of the similarity weight. Similarity Weight calculation is represented by the Equation (1), where Pr is the probability and λ is the hyper parameter.

$$SimilarityWeight = \frac{\sum((Residual)^2)}{\sum(P_r(1-P_r)+\lambda)} \tag{1}$$

The information gain is obtained through the difference between the sum of the similarity weight of both the leaves and the similarity weight of Root Node. The sum of the sigmoid function of the base learner's prediction and the learning rate used to determine the decision tree is the predicted probability. The difference between the prior residual and the predicted probability can be used to calculate residual prediction. On continuation of the iteration for the remaining input categories, the minimal residual could be obtained at the end of the iteration, implying that the XGBoost had been properly trained. Furthermore, the non-significant branches of a decision tree could be eliminated if the decision tree has a

large depth. This method is known as backward pruning or post pruning. The tree does not need to be divided further if the information gain is smaller than the cover value, i.e., Pr(1-Pr). When we post-prune a decision tree, we commence by constructing the (whole) tree and then adjusting it to improve the accuracy on unknown occurrences. The XGBoost classifier's classified instances, as well as the top 20 % of the metadata obtained in the previous step, are used as inputs to compute semantic similarity using the NPMI measure (normalized pointwise mutual information measure) and the Jaccard similarity measure. It is concerned with individual occurrences, unlike mutual information, which is based on PMI, whereas MI is concerned with the average of all possible events.

$$npmi(x;y) = \frac{pmi(x;y)}{h(x,y)} \tag{2}$$

Equation [2], depicts the Normalized Pointwise Mutual Information where x and y are a pair of outcomes belonging to discrete random variables that can be normalized between [-1,+1], with -1 (in the limit) indicating that they never eventuate together, 0 indicating independence, and +1 indicating complete co-occurrence. h(x) is the fundamental quantity generated from a random variable's likelihood of a specific event occurring. Similarly, Jaccard Similarity is represented by the Equation [3], where J stands for Jaccard distance. As illustrated below, the size of the intersection is divided by the size of the union of two sets, A and B.

$$J(A,B) = \frac{|(A \cap B)|}{|A \cup B|} \tag{3}$$

In this scenario, the NPMI threshold is considered as 0.5, and only positive NPMI values between zero and one are taken into account. However, for Jaccard similarity 0.75 is considered. Hence these are the two objective functions. Semantics in the measures, namely NPMI and Jaccard, are used as objective functions for the charge system search algorithm. A meta heuristic optimization algorithm is charge system search. They're utilized to refine the best solution from the initially obtained solutions. The number of initial cluster centers is set, and the dataset(K) is loaded as the first step followed by the initialization of Charged Particles' (CP) locations and velocities. The mass for each charged particle out of the total number of cluster centers and for each feature are determined in the dataset using equations [4] and [5].

$$C_k = X_(min,i) + r_i * (X_(i,max) - X_(i,min)) \tag{4}$$

where i=1,2…,n and k=1,2...,K

$$m_k = \frac{fit(k) - fit(worst)}{fit(best) - fit(worst)} \tag{5}$$

Using the sum of squared distances, the value objective function is evaluated and the items to the clusters with the lowest objective function value is allocated. A new variable called CM is used to store the positions of initial charged particles (C_k). The value of moving probability for each charged particle Ck is determined, and the value of separation distance is calculated. Using the equations [6] and [7] the new positions and velocities of charged particles is calculated, where random functions are depicted by rand1 and rand2, whose values lie in between 0 and 1. The control parameters are depicted by Z_a and Z_v which regulates the actual electric force and past velocities, m_k is the mass of kth CPs which is equal to the q_k and Δ represents the time step that is assigned as 1.

$$C_(k,new) = rand_1 \times Z_a \times \frac{F_k}{m_k} * \Delta t^2 + rand_2 \times Z_v * V_(k,old) \times \Delta t + C_(k,old) \tag{6}$$

$$V_(k,new) = \frac{C_(k,new) - C_(k,old)}{\Delta t} \tag{7}$$

The newly created charge particles are compared to the value of the objective function to the Charge particles living in CM after recalculating the value of the objective function using updated placements of charged particles. With the help of the memory of all previous solutions achieved in the iterations, the optimal solution is found. The initially obtained solution is the semantically similar entities computed using NPMI and Jaccard. Furthermore, the charge system search refines the initial solution by applying the same objective functions based on the charge system search criterion. The instances are arranged in the increasing order of the NPMI Measure to finally rearrange. It is ranked in the increasing order of the NPMI measure, and the results are yielded to the user. Results are the indicative facets along with the microblogs themselves. The suggested terms in the microblog that are the microblogs' keywords are known as indicative facets. Along with the microblog, the categories and keywords are retrieved. If the user is satisfied, the search ends; otherwise, the current user clicks, which are based on the blogs user clicks, or the facets that were advised, or the clicks that were recorded, are input into the semantic similarity computation scheme in the second parse. This process is repeated until the user is satisfied with the results. When no further user clicks are recorded, the microblog's recommendation comes to an end.

V. Performance evaluation and results

TABLE I: COMPARISON OF PERFORMANCE OF THE PROPOSED SAMBT FRAMEWORK WITH OTHER APPROACHES

Model	Precision	Recall	Accuracy	F-measure	FDR
TRFLWE	84.38	86.78	85.58	85.56	0.16
WLSTM	90.11	92.15	91.13	91.11	0.10
ITRM	91.47	93.67	92.57	92.55	0.09
K-Means + Clustering Collaborative Filtering	82.37	84.37	83.37	83.35	0.18
Proposed SAMBT	95.49	97.71	96.60	96.58	0.05

From Table 1, it is indicative that TRFLWE 84.38 % of precision, 86.78 % of recall, 85.58 % of accuracy, 85.56 % of F-Measure with an FDR of 0.16. Similarly, the WLSTM is 90.11 % of precision, 92.15 % of recall, 91.13 % of accuracy, 91.11 % of F-Measure with an FDR of 0.10, ITRM yields 91.47 % of precision, 93.67 % of recall, 92.57 % of accuracy, 92.55 % of F-Measure with a FDR of 0.09. Similarly, the combination of K-Means clustering with collaborative filtering churns 82.37 % of precision, 84.37 % of recall, 83.37 % of accuracy, 83.35 % of F-Measure with an FDR of 0.18. However, the suggested SAMBT framework has the highest average precision of 95.49, highest average recall of 97.71 %, highest average accuracy of 96.60 %, highest average F-measure of 96.58 %, with the lowest FDR of 0.05. The suggested SAMBT is a microblog tag recommendation model that uses a bi-classification model, which explains why it has the highest precision, recall, accuracy, and F-measure, as well as the lowest FDR measure. The model initiates meta data generation, which is classified using the deep learning model, WLSTM. The purpose for generating a classification of meta data is to increase the amount of cognitive local knowledge which is included into the model. The dataset, on the other hand, is classified using the XGBoost classifier, a machine learning classifier in which the feature selection is not done automatically. The XGBoost classifier is not a deep learning classifier because the feature selection had to be non-auto hand crafted in order for the XGBoost model to fit into the dataset and be classified based on the manually picked features to ensure explainability. However, using both the XGBoost algorithm and the WLSTM model to classify the dataset, which is relatively shallow when compared to the meta data, which is a deep learning model, ensures a high rate of success, increasing the model's classification power. The inclusion of NPMI with jaccard similarity with varying thresholds for computing semantic similarity ensures that the relevance is very high when compared to the baseline model, and the meta heuristic algorithm, charge system search ensures the computation of the most optimal solution sets from the initial set of feasible solutions, resulting in better optimal final tags that are finalized based on the model. The integration of all of these entities, including deep learning, machine learning models, charge system search, NPMI, and jaccard similarity, assures that the model outperforms the baseline model. The SAMBT model performs better at recommending tags for microblogs. As a result, the suggested SAMBT surpasses the baseline models. TRFLWE has the second lowest precision, recall, accuracy, and f-measure, as well as the second highest FDR value, due to the fact that it is based on the relationship analysis between the words in the text. Word embedding is combined with feature learning and a skip-gram model. The Skip Gram model is a traditional paradigm that involves feature and relationship learning. However, there is no auxiliary knowledge which is fed into the model. The word embeddings alone serve as the only background knowledge. Furthermore, the knowledge that has been absorbed, as well as the relevance computational paradigm, is simply a learning algorithm

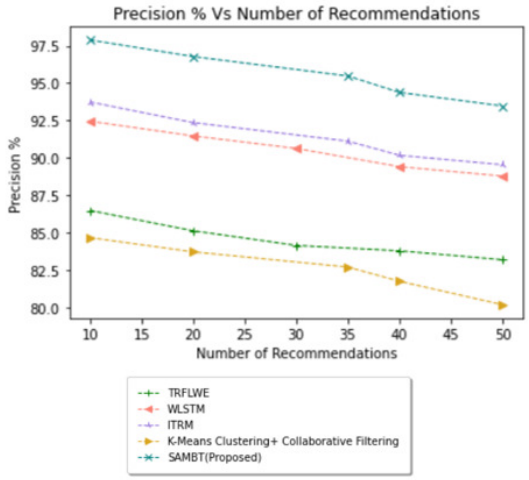

Fig. 2: Precision percentage vs Number of Recommendations

that causes it to lag. The WLSTM model is a sequence-to-sequence model based on an attention-based encoder and an LSTM encoder. It ensures multi-label classification with global relationship learning. Also, the model incorporates local positional encoding. Local positional encoding is also included in the model. Hence, all these models are sequence to sequence models that rely solely on LSTM global linkages learned from the environment and data from the world wide web. However, there is no clarity on how to maximize the quantity of global knowledge that should be incorporated into the model. As a result, the model will have a lot of overfitting and hence will cause a lag. In ITRM, tag correlation with content tag overlap is staffed, which is also model for textual content. The encoder uses RNN, while the decoder deals with tag connection by constructing a prediction path, resulting in tag content overlap. Even this approach, however, falls short because there is a gap in the learning and optimization of global knowledge. The relevance computational model is based on the encoder and decoder stages of a sequence-to-sequence model. Specific regulatory mechanisms, on the other hand, are missing, and thus model ITRM suffers lags. The K-means clustering with the collaborative filtering yields the least amount of precision, recall, accuracy, and the highest FDR value because K-means clustering is a clustering algorithm and for the collaborative filtering model, all the tags had to be retread and item rending computation model has to take place. In reality, every entity on the world wide web recommended tags cannot be treated as results, this model also fails abruptly. Fig. 2. depicts the precision vs number of recommendations distribution curve for the proposed recommendation model for microblog tag recommendation for the baseline model. It is very clear that the curve of the proposed SAMBT has the highest precision vs number of recommendations when compared to the baseline models irrespective of the number of recommendations. As a result, the suggested hybridized semantic model SAMBT has higher precision, recall, and accuracy than the baseline models.

VI. Conclusion

Only the top 20 % of categorized metadata instances are processed in the proposed semantically aware micro-blog tag recommender, encompassing bi-classification model. The suggested model was able to achieve the F-measure with the lowest FDR value using techniques such as LSTM classifier, charged system search, and XGBoost classifier using NPMI and jaccard similarity. When charting the graph to view the precision data against the number of recommendations, it is clear that SAMBT outperforms the baseline models.

References

[1] Tang, S., Yao, Y., Zhang, S., Xu, F., Gu, T., Tong, H., ... Lu, J. (2019, July). An integral tag recommendation model for textual content. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 5109-5116)

[2] Zheng, L., Tianlong, Z., Huijian, H., Caiming, Z. (2020). Personalized tag recommendation based on convolution feature and weighted random walk. International Journal of Computational Intelligence Systems, 13(1), 24-35.

[3] Lei, K., Fu, Q., Yang, M., Liang, Y. (2020). Tag recommendation by text classification with attention-based capsule network. Neurocomputing, 391, 65-73.

[4] Najafabadi, M. K., Nair, M. B., Mohamed, A. (2021, January). Tag recommendation model using feature learning via word embedding. In 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI) (pp. 000305-000310). IEEE.

[5] Fletcher, K. K. (2020, June). An Attention Model for Mashup Tag Recommendation. In International Conference on Services Computing (pp. 50-64). Springer, Cham.

[6] Sun, J., Zhu, M., Jiang, Y., Liu, Y., Wu, L. (2021). Hierarchical attention model for personalized tag recommendation. Journal of the Association for Information Science and Technology, 72(2), 173-189.

[7] Yang, Z., Lin, Z. (2021). Interpretable video tag recommendation with multimedia deep learning framework. Internet Research.

[8] Zhao, W., Shang, L., Yu, Y., Zhang, L., Wang, C., Chen, J. (2021). Personalized tag recommendation via denoising auto-encoder. World Wide Web, 1-20.

[9] Roopak, N., Deepak, G. (2021, March). KnowGen: A Knowledge Generation Approach for Tag Recommendation Using Ontology and Honey Bee Algorithm. In European, Asian, Middle Eastern, North African Conference on Management Information Systems (pp. 345-357). Springer, Cham.

[10] Liu, S., Liu, B. (2021, October). Personalized Social Image Tag Recommendation Algorithm Based on Tensor Decomposition. In 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC) (pp. 1025-1028). IEEE.