Deep Learning for Latent Events Forecasting in Content Caching Networks

Zhong Yang, *Student Member, IEEE*, Yuanwei Liu, *Senior Member, IEEE*, Yue Chen, *Senior Member, IEEE*, and Joey Tianyi Zhou, *Senior Member, IEEE*

***Abstract*— A novel Twitter context aided content caching (TAC) framework is proposed for enhancing the caching efficiency by taking advantage of the legibility and massive volume of Twitter data. For the purpose of promoting the caching efficiency, three machine learning models are proposed to predict latent events and events popularity, utilizing collected Twitter data with geo-tags and geographic information of the adjacent base stations (BSs). Firstly, we propose a latent Dirichlet allocation (LDA) model for latent events forecasting because of the superiority of LDA model in natural language processing (NLP). Then, we conceive long short-term memory (LSTM) with skip-gram embedding approach and LSTM with continuous skip-gram-Geo-aware embedding approach for the events popularity forecasting. Furthermore, we associate the predict latent events and the popularity of the events with the caching strategy. Lastly, we propose a non-orthogonal multiple access (NOMA) based content transmission scheme. Extensive practical experiments demonstrate that: 1) the proposed TAC framework outperforms conventional caching framework and is capable of being employed in practical applications thanks to the associating ability with public interests; 2) the proposed LDA approach conserves superiority for natural language processing (NLP) in Twitter data; 3) the perplexity of the proposed skip-gram based LSTM is lower compared with conventional LDA approach; and 4) evaluation of the model demonstrates that the hit rates of tweets of the model vary from 50% to 65% and the hit rate of the caching contents is up to approximately 75% with smaller caching space compared to conventional algorithms. Simulation results also shows that the proposed NOMA-enabled caching scheme outperforms conventional least frequently used (LFU) scheme by 25%.**

***Index Terms*— Machine learning (ML), supervised learning, neural networks, edge computing.**

I. Introduction

RECENT advances in mobile smart devices, ubiquitous social media and application brings tremendous expansion of mobile data traffic. The visual networking index (VNI) report from Cisco [2] reveals that global mobile data traffic (GMDT) is expected to grow to 77 exabytes per month by 2022, a seven-fold increase over 2017. Specifically, 5G will be 3.4 percent of connections but 11.8 percent of total traffic by 2022. The extensive GMDT growth and innovative network technology compel network providers to investigate novel techniques for sufficing the network services and easing up backhaul transmission.

Content caching at networks edges is a prospective approach for reducing the network backhaul transmission and bringing down the network delay [3], [4]. Nevertheless, caching superiority is related closely with the popularity of the contents in the network. In conventional caching frameworks, the user preference is assumed following a generalized Zipf law [5]: the content request rate $r(i)$ for the i-th content in the network is proportional to $1/i^{\alpha}$ where α is the temperature of the network and typically less than Unit. However, the Zipf law is an experience distribution and lacks theoretical guarantee. Specifically, the counting methods based on Zipf distribution properties only demonstrate the frequency of the words rather than the concurrency between the words. Therefore, the extracted events are incomplete, which impairs the performance of text-related content prediction in wireless caching. Moreover, employing the properties of Zipf law of text-related content to determine what to cache regardless of the locations of base stations (BS) is pervasive yet prodigal [6], especially for the legibility and massive volume of social media data. According to [7], social media usage is one of the most popular online activities and the number of people using social media worldwide increase to almost 3.43 billion, 3.5 times that of 2010.

Social media motivated caching strategy has attracted attentions from both academia and industry [8]–[11]. The authors in [8] analyse the Twitter data of 2016 U.S. presidential election utilizing LSTM networks to reduce the service latency. A preference-aware optimization [9] considers user side adaptive streaming, coordinated bandwidth allocation, and network side caching content selection. In [10], caching cost of the base stations (BS) and social factors among mobile users (MU) are considered in ultra-dense small cell networks (UDN) to obtain effective caching incentives and the optimal social group utility. Zeydan, etal, [11] proposed a big data enabled caching architecture, in which a vast amount of data is harnessed for content popularity estimation and content caching. Twitter is one of the most popular social media platforms in the world, that contains countless open accessed tweets (messages) published by the users from different regions. With billions of new tweets being posted every day, the freshness and chronological variation of the text contents in tweets are attracting more and more researches to

Manuscript received September 11, 2020; revised February 2, 2021 and May 10, 2021; accepted June 27, 2021. Date of publication July 26, 2021; date of current version January 10, 2022. This article was presented in part at the IEEE Global Communication Conference (GLOBECOM) 2019 [1]. The associate editor coordinating the review of this article and approving it for publication was C. Huang. *(Corresponding author: Zhong Yang.)*

Zhong Yang, Yuanwei Liu, and Yue Chen are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: zhong.yang@qmul.ac.uk; yuanwei.liu@qmul.ac.uk; yue.chen@qmul.ac.uk).

Joey Tianyi Zhou is with the Agency for Science Technology and Research (A*STAR), Singapore 138632 (e-mail: joey.tianyi.zhou@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2021.3096747.

Digital Object Identifier 10.1109/TWC.2021.3096747

exploit tweets to gather large amounts of public data for big data related researches.

As the public tends to post their preferences and interests on social media platforms, it brings us an opportunity to cache text-related content more accurately in the BS through topics/events prediction. To efficiently predict caching text-related contents among different BS, it is plausible to associate the public preference with the Twitter topical issues. Top words in Twitter topics have been proved to vary according to different locations in London [12]. Twitter data indicate the preference of the public towards specific topics [13]. After extracting latent events from tweets, the text-related contents caching in the BS can be determined. However, the Zipf-distribution is not sufficient accurate since the structure of the natural language shows a statistic structure beyond Zipf distribution properties [14], which leads to incapability of Zipf-based algorithms in extracting latent events from tweets..

Big data techniques have enabled the industry to deal with large amounts of data in high efficiency. Wireless content caching system is the crucial key to improve the efficiency of the wireless caching at the edge of the wireless network. However, it is always nontrivial to determine what to cache at the wireless devices [15]. With the aid of friendly application programming interface (API) offered by the social media platforms, we are able to easily filter the social media contents of large amounts within the constrained region. Since the APIs also enable users to filter the regions of tweets based on their location tags (geo-tags), relating regional preference to the tweets in that region seems to be a feasible way to determine what to cache based on the public preferences. Combining the APIs with big data techniques, individual BS is therefore capable of automatically allocate the local public-posted information. Mentioned above, with the aid of the machine learning (ML) approaches, the preference is predicted and the text-related caching contents are associated with the preference of the regional public.

A. Prior Work

Caching at the edge of the wireless networks is considered one of the most important direction due to its great potential of forwarding desired contents during the rush hours, which alleviates the network burden [16]. With the increased number of mobile users in wireless networks, pushing frequently requested contents close to them has been deemed as an efficient way of overcoming the bottlenecks of the wireless communication systems. Furthermore, caching contents according to the preference of the MUs can improve the efficiency and prevent congestions to some extent [15]. According to [17], video content caching approaches have been proved to be efficient in improving video throughput.

ML approaches and social media platforms have attracted increasing more attention in the field of wireless caching recently. With the aid of ML approaches, proactive allocating systems are able to better predict the wireless traffic patterns [18]. By exploiting topical issues in the regions and the interactions among the public, the networks are capable of better predicting the text-related contents at the edge of the networks. To reinforce the quality of the wireless network service, caching has been proved to be effective in caching radio contents [19] and the hit rate has been improved.

Besides, with the aim to establish a proactive device-to-device (D2D) caching network, authors of a paper linked the users together to share caching contents [18]. Aiming to extract contextual contents from users' interactions, there is also a paper proposed a framework of wireless caching [20]. According to the prior works, the application of ML approach has been applied on the video caching field to reinforce the wireless caching network [19]. There is also a paper which employs several parameters to evaluate the different candidate contents through making evaluations on their popularity profile [21].

Related works include the effort to associate different users of the social networks with the aim to establish a proactive D2D caching network [18]. The thought is to retrieve the cached contents from other users to satisfy the requests from others users for the same contents. Similar to this work, there is also a paper proposed a framework of wireless caching [20], which extracting the contextual information from the users' interactions. The results mentioned in the above two paper demonstrate that the contextual extraction wireless caching models are capable of reinforcing the accuracy of wireless caching and reducing the redundancy of the BS caching contents. Acknowledged from the prior works, the application of ML approach has been applied on the video caching field [19]. In this paper, the authors demonstrated several approaches towards determining the video contents to be cached at the BS, including Least Recently Used (LRU), Least Frequently Used (LFU) and their proposed ML approach. The hit rates of their models vary from 80% to 90% as the caching size varies from 10 GB to 100 GB. However, in their model, the monotonic video caching dataset is exploited, which related poorly to the text and other aspects of public preference. There is a paper exploits several parameters to determine the popularity of the candidate contents to be cached [21]. In this paper, authors classify the contents based on their "popularity profile", which is determined by algorithms. After the evaluation, the data with high popularity profile are cached at the BS while the other are downloaded from the network. Final result demonstrated that the fetch rate increases along with the popularity profile while the fetch rates vary from 15% to 25% based on their proposed algorithm.

Therefore, based on the two prior papers mentioned above, we are capable of concluding that the prediction and determination of complex, comprehensive wireless caching contents, which including different types of data (video, images, text, etc.) are increasingly more nontrivial compared to the wireless caching model towards one monotonic caching content. Further, deep learning approaches have been proved to be effective and efficient while determining wireless caching contents.

B. Motivation and Contribution

There are some research contributions related with this topic, such as context-aware schemes [22], [23] and content

TABLE I
COMPARISON BETWEEN THIS PAPER WITH EXISTING ML METHODS

	Proposed caching scheme	[8]	[22]	[23]	[24]	[25]
Content caching	✓	✓	✓	✓	✓	✓
Content distribution forecasting	✓		✓	✓	✓	
Semantic analysis	✓	✓				
NOMA-aided delivery	✓					
OMA-aided delivery	✓					
Practical experiments	✓	✓		✓		
ML solutions	✓	✓	✓	✓	✓	✓

The marks mean that the contents in the leftmost column are considered in the corresponding paper.

recommendation schemes [24], [25]. The authors in [22] proposed a posteriori caching scheme that decides the content placement before the analysis of the placement gain. In [23], a context-aware caching scheme is proposed, which learns the content popularity in an online manner. However, different from references [22], [23], our work consider the twitter aided content caching framework, because twitter data is capable of efficiently indicating the preference of the public towards specific topics. Although the data set from MovieLens is used in [22], the semantic environment of twitter data is more complex than that of MovieLens. The authors in [24], [25] proposed cache-aware recommendation systems, in which the recommendation strategy is adopted to enhance the gain of content caching. However, the proposed recommendation systems require the users' preference as prior information, which is often hard to obtain. Therefore, the proposed ML based latent events forecasting can work as a complementary for references [24], [25]. The predicted latent events can be used for the recommendation systems in [24], [25] to improve the performance of the recommendation policies. The comparison between this paper and previous works is given in Table I. Noting that references [23]–[25] only considered content placement in caching networks, content delivery is not considered, therefore, [23]–[25] cannot be extended to NOMA/OMA case. Meanwhile, the transmission delay is analysed in references [8] and [22], therefore [8] and [22] can be extended to NOMA/OMA case.

Although previous research contributions have investigated social media aided content caching. However, the practicality of ML based twitter data analysis for wireless caching is still in infancy, especially for novel recurrent neural networks, such as long short-term memory (LSTM) networks with skip-gram embedding. Moreover, previous works mainly use conventional orthogonal multiple access (OMA) schemes for content delivery. In this work, we conceive a non-orthogonal multiple access (NOMA) based content transmission, which is capable of transmit multiple contents simultaneously, thus reducing the delay of content delivery [26], [27].

Our motivation is to combine different types of ML models to propose a wireless caching framework, which exploits social media data as reference to determine the text-related contents at BS. Since there are different data types in the twitters including images, videos, music, etc., the proposed model is general and feasible for multiple caching context. As mentioned above, social media platforms have been regarded as a major network traffic consumer due to their popularity [28]. After determining what heavy-weighted caching contents (images, videos) to cache, the caching-enabled wireless BS is capable of reducing the burden of the network and reduce backhaul capacity. Since allocating different events of text-related caching content requires distinct parameters of Zipf-based models in different areas to obtain satisfying caching predictions, the efficiency of setting up such networks is limited and the interactions among the networks are limited. Therefore, an autonomous and reliable wireless caching framework is desired. Furthermore, when indigenous BSs are capable of forming an autonomous region to reinforce multi-BS caching, which leads to less computing costs and wastes.

One core problem is that tweets are considered to be challenging for events allocating due to their colloquialism, short length (less than 280 components) as well as the informal usage of language [29]. The conventional counting approaches, like latent semantic analysis (LSA) [30], are insufficient to find the events. To enhance the accuracy, we apply two ML models, namely latent Dirichlet allocation (LDA), and long short-term memory (LSTM). The Beyes topic modelling approach, latent Dirichlet allocation (LDA), which employ multinomial probability over terms for topic allocation [31]. LDA is able to obtain precise outcome from data of social media like Twitter [32]. As the preference of the public in the same region tends to vary chronologically, we employ long short-term memory (LSTM) [33] to model long-term contextual information. For short-sequence time series prediction problems, a conventional recurrent neural network (RNN) algorithm is efficient. However, in our proposed twitter based latent events prediction framework, the input sequences and output sequences are long, which makes conventional RNNs insufficient. The LSTM networks have forget gate, input gate, and output gate, which work to identify and filter useful information. Therefore, LSTM networks are capable of fitting the complex relationship between history twitter data and future twitter data, and are more suitable for the proposed twitter based latent events prediction framework compared with conventional RNNs. [34].

The other core problem is how to associate the tweet text, which represents the public preference, with the solid caching contents. In an effort to associate Twitter events with corresponding BS, we propose the approach to arrange the tweets to their pertinent BS in London. Previous studies on Twitter demonstrate that Twitter events with geographic information represent the preference of the public. After obtaining the events from the tweets, we associate the events with actual tweet text to determine what media files to cache in order to achieve better performance. Considering the difference between that the caching background of social media platforms and that of traditional content streams (i.e. the video sites), we propose criteria to demonstrate the coherency between the events and future media contents. Our contributions are summarized as follows:

- We propose a novel Twitter context aided content caching (TAC) framework in which latent events are extracted from collected Twitter context with BS geography information utilizing a versatile LDA model, due to its superiority for natural language processing (NLP). Moreover, deep learning approaches are developed for reinforcing the autonomous association of the predicted latent events and wireless text-related media files.
- We develop LSTM networks with skip-gram-Geo-aware embedding that is capable of predicting not only the terms of the Twitter events but also the location where the tweet are posted, which conventional LSTM networks cannot.
- Extensive practical experiments demonstrate that: 1) Hit rates of ML based TAC approach outperforms LFU and LRU by approximately 12% and 35%, respectively. 2) Hit cache portion of ML approach is close to that of LFU, and outperforms LRU by around 33%. 3) The proposed NOMA-enabled caching scheme outperforms NOMA-LFU scheme by 25%

To make the following sections clear, some terms are defined as follows:

- *Event/Topics:* The extracted information from the twitter data.
- *Documents:* The collected twitter data in different locations (The area of London is divided into nine areas according to latitudes and longitudes of BS locations).
- *Words:* Collected tweets that is used as inputs for the proposed ML algorithms. To satisfy the prerequisites of training the proposed ML models with the collected tweets, some characters are deleted based on the standards in section IV-A.
- *Contents:* The associated media data with the collected tweets, which are normally images or videos. They are capable of generating remarkable impact on the network backhaul loads.

C. Organization

The paper is arranged into following sections. In Section II, related works, which employ ML approaches to solve wireless caching problems, as well as their performances are listed. In Section III, the structure of the framework is demonstrated. In Section IV, the problems of events extraction and prediction along with the solutions are demonstrated. In Section V, experiments are proposed to evaluate the model. Numerical and literal results are listed to evaluate the events extraction procedures. In Section VI, the events obtained from Section V are applied to determine the wireless caching contents. Certain numerical results are demonstrated to present the accuracy and efficiency of our approaches. Table I provides the comparison between this paper and previous works.

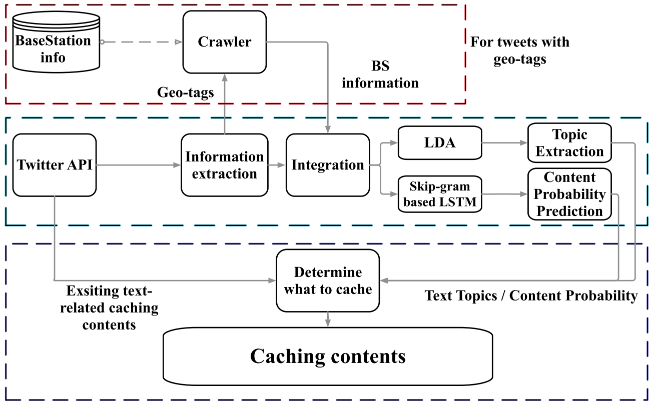


Fig. 1. The proposed Twitter aided content caching framework.

II. TAC Framework

In this section, the proposed TAC framework and process of Twitter dataset preparation are demonstrated.

The TAC framework is illustrated in Fig.1. The tweets are collected through the Twitter API with their geo-tags. After extracting the geo-tags from the tweets, the information of the nearest BS is obtained through a python Crawler. After offering the crawler with the geo-tags from tweets, the crawler returns the locations of BS.[1] Afterwards, the dataset is arranged into different regions as an integration of the text with the geographic information. Thereinafter, three ML models are invoked to generate predictions according to the collected dataset from Twitter. Finally, since the Twitter API is offering URLs directing to the media files (images, videos, etc.) if the files are contained in a tweet, the framework determines which of the files to cache based on the descriptions in tweets associated with the events forecasting. The dataset throughout the paper was collected through the Twitter API with filter, which gathers the tweets with geo-tags in London. The latitude and the longitude restrictions are given as follows:

- Latitude restriction from 51.7136401 to 51.3679144
- Longitude restriction from 0.285472 to -0.4488468

The proposed framework for latent events prediction contains three steps: Step 1: The latent events are predicted using the collected twitter data; Step 2: The predicted context is associated with other tweets that contains the media data (video or image). Step 3: the corresponding media data (video or image) is cached to the pertinent base stations (BSs), which improves the caching performance, such as improving the hit rate and reducing the contents transmission delay.

III. LDA Learning for Latent Events Forecasting

After demonstrating the structure of the proposed TAC framework, this section illustrate text-related caching contents based on twitter events. Regarding our proposed TAC

[1] BS data retrieved from https://mastdata.com

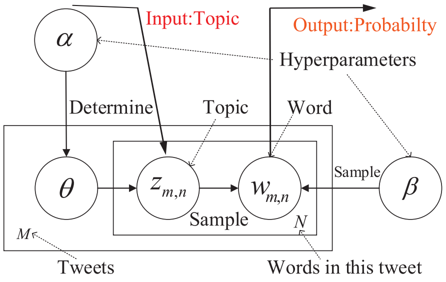


Fig. 2. The structure of LDA model for latent topic forecasting.

framework, obtaining Twitter events is necessary for determining the caching contents. In this section, we use novel ML approaches to extract and predict the events from the collected tweets. Firstly, we illustrate how the LDA approach is exploited to extract latent events from the existing tweets. Then, we demonstrate how to chronologically predict future tweets based on LSTM. In the end, we display how to exploit our new embedding method to chronologically predict future tweets associated with geographical locations based on LSTM.

A. Extracting Latent Events From Tweets

The introduction section has demonstrated that the twitter events are associated strongly with the public preference in prior works, in which approach are we able to extract the events with existing twitter text is a crucial term towards realising our proposed wireless caching framework. In this section, we demonstrate how to exploit LDA method to obtain satisfying events extraction.

1) The LDA Model: LDA is an unsupervised Beyesian probabilistic model with the objective to identify the probable topics from the documents [35]. The proposed LDA model for latent events forecasting is illustrated in Fig. 2. As in Fig. 2, the idea is to determine the corresponding topic $z_{m,n}$ of the n^{th} word in m^{th} document based on parameter α while to determine the probability $w_{m,n}$ of each word under the given parameter β and a given topic in order to obtain every word in this given topic. Each document is assigned a distribution to the latent events. Through setting the K latent events, the words are associated with the events as $P(events|document)$ and $P(events|term)$ [32].

The objective is to obtain the probabilities that each word belongs to different latent events, and therefore, we allocate events by listing the words of the highest probability in each latent event, illustrated in Eq. (1),

$$P(Z|W,\alpha,\beta), \tag{1}$$

where Z is the event, W stands for the document, and α, β are parameters.

According to Fig.2 and its approach originated from [35], the probability equation is given by Eq. (2),

$$p(\theta|\alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} (\prod\nolimits_i \theta_i^{\alpha_i - 1}), \tag{2}$$

where θ is the event multinomial distribution parameter, α is a k-dimensional vector of the Dirichlet districution. θ complies to the Dirichlet distribution of parameter α.

Algorithm 1 LDA Learning for Latent Events Forecasting

```
For every word $w$ in each document, assign the word to a random latent event $z$;
repeat
    Scan the corpus, for every word $w$ in the corpus, do Gibbs Sampling, and obtain its latest latent event;
    update the corpus;
until The Gibbs Sampling Converges OR Reach the maximum iteration number
Gather the event-word ($W$-$Z$) Co-occurrence Matrix, which is the result of LDA model;
Obtain the event distribution of the model $\vec{\theta}_{new}$;
```

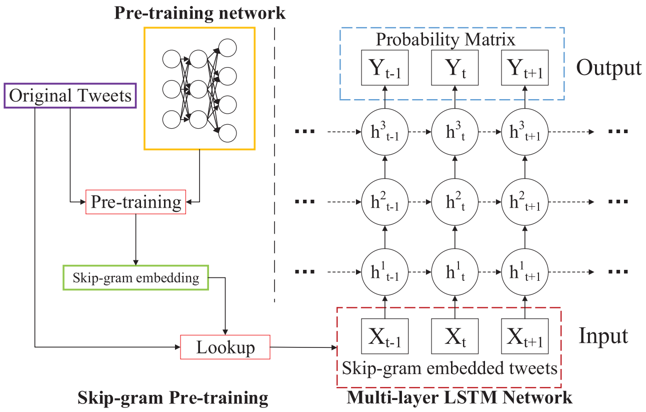


Fig. 3. LSTM model for content popularity forecasting.

With given parameter α and β, the joint probability distribution is able to be demonstrated in Eq. (3),

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod\nolimits_{n=1}^{N} p(z_n|\theta) \cdot p(w_n|z_n, \beta), \tag{3}$$

where z, w stand for event set and document set, N stands for the N words in document w. As θ, z are latent variables, we adopt Gibbs Sampling to marginalize them.

2) Model Learning: The learning procedure of the LDA model is based on the Gibbs Sampling. The learning procedures of the LDA model are demonstrated in **Algorithm 1**. Here, we set the quantity of iterations to 100, the size of the latent events to 20, and words per event to 7.

B. Forecasting Events Based on Tweets

LSTM is compatible for chronological data forecasting. Compared with conventional RNN, the LSTM model has the advantage in memorizing the long-term memory [36]. The structure of the proposed LSTM model is illustrated in Fig. 3. In Fig. 3, LSTM cells are basic units of the LSTM model. The two arrows surrounding the cell $Cell_t$ are the vector transfer of the long-term memory c_{t-1} from the former cell $Cell_{t-1}$ and the short-term memory h_{t-1} from the former cell $Cell_{t-1}$. Inside the cell are neural network Layers, including tanh layers and sigmoid layers. X_t and h_t are the input and output at the time t. The algorithm flow of training the LSTM model is listed in **Algorithm 2**

Algorithm 2 LSTM for Content Popularity Forecasting

```
Initial the model layers and hidden sizes with given parameters;
Feed the first batch of input $X_1$, calculate the memory $h_1$ and Cell State $C_1$;
repeat
    For time stamp $t$, the forget gate layer $\sigma$ determine what parts of the cell state $C_{t-1}$ are preserved;
    For time stamp $t$, the input gate layer $\sigma$ determines which value between $h_{t-1}$ and $x_t$ is updated. And derive the cell state of this time stamp $C_t$, memory $h_t$ and candidate value $\widetilde{C}_t$ through $tanh$ layers and operators;
    Invoke a sigmoid layer (output gate layer) to determine what parts of the cell state are the output and update the cell state;
    Evaluate perplexity through calculating cross entropy between the prediction and $X_{t+1}$;
    Invoke the gradient descent optimizer with a given learning rate to optimize the model;
until Reach the maximum epoch number
Save the model;
```

TABLE II
PARAMETERS OF LSTM MODELS

Parameter	*Medium*	*Large*
Initial scale of weights	0.04	0.05
Initial learning rate	0.1	0.2
maximum permissible norm of the gradient	5	10
number of LSTM layers	2	3
number of unrolled steps of LSTM	20	50
hidden size	650	1500
max epoch	65	55
learning rate decline epoch	25	30
learning rate decline rate	0.8	2/3

The skip-gram embedding approach is commonly employed in natural language processing (NLP). The idea is to feed the model with the words of adjacent positions and therefore associate the words with the context. For example, while skip is 1 and a batch is n + 1 words, we firstly feed the model with the range of i^{th} to $(i+n)^{th}$ words of the text, and the second step is to feed the $(i+1)^{th}$ to $(i+n+1)^{th}$ words [37].

1) Model Training: To train the LSTM model, we exploited the Twitter dataset from 27th January 2018 to 26th February 2018 (30 days) as the training dataset, and tweets of 27th February 2018 as the testing dataset. Here, the dictionary is limited to be 60000 words. To fulfill the prerequisites, each tweet is mapped into a vector based on the indexes of each word in the dictionary. The vectors are fed into the network based on their chronological order. The parameters of the networks are detailed in Table II. We use a gradient descent optimizer in TensorFlow to automatically adjust the learning rate.

2) Forecasting: The model predicts the terms of the events through a softmax (normalized exponential) layer, which is commonly used loss function for multi-class classification [38]. The events are sorted according to the predicted probabilities and the events with higher probabilities are cached in our framework.

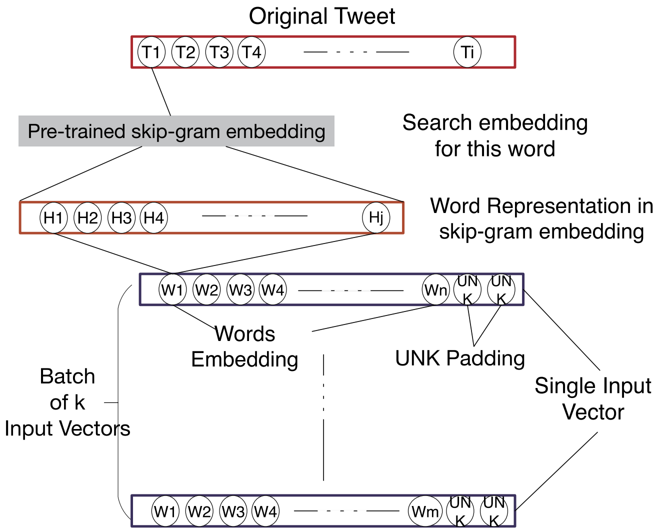


Fig. 4. LSTM inputs based on skip-gram embedding.

C. Predicting Events Based on Tweets and Geographical Locations

The anticipation of this model is to combine the geographic information with the Twitter text. To achieve the objective, we append the tokens indicating the latitude ranges and longitude ranges to the tweet vectors. By feeding the LSTM model with the embedded vectors including the tokens, the text are therefore able to be associated with the geographic information.

1) Model Training: In this new model, we apply BoW embedding to map each tweet into a vector. Regardless of the order of the components, we merely assume the terms are ingredients of the events of the tweet. The skip-gram approach is invoked to feed the vector of tweets into the model. The process is illustrated in Fig. 4. Fig. 4 presents the inputs of the proposed LSTM for latent events prediction. In Fig. 4, each single tweet is mapped into a vector based on its content and the geographic information of its BS. Thus, for the first n elements (W_1 to W_n) in single tweet vector, they are the numerical representation of the words in this tweet based on the index in the above dictionary.

After the data cleaning procedure mentioned in data preparation section (removing the meaningless parts in tweet text), most tweets are able to be restrained within 20 words (approximately 97%). While larger vector leads to lower accuracy, we constrain the vector size to 20 elements. For those less than 20 words, UNK tokens are filled into the vector. For tweets longer than 20 words, we truncate the first 20 words where the tags and most words are preserved. Afterwards, the last two elements in the tweet vector represent the latitude range and the longitude range of the tweet. The Latitude token ranges from 81112 to 81114 while the Longitude token ranges from 91112 to 91114 (to avoid the collision with the indexes of the dictionary). Afterwards, every m tweet vectors are arranged

TABLE III
PARAMETERS OF LSTM MODELS

Parameter	***Medium***	***Large***
Initial scale of weights	0.04	0.05
Initial learning rate	0.1	0.2
maximum permissible norm of the gradient	5	10
number of LSTM layers	2	3
number of unrolled steps of LSTM	50	100
hidden size	650	1000
max epoch	45	55
batch size	20	20
learning rate decline epoch	25	30
learning rate decline rate	0.8	2/3

into the single input vector based on their chronological order. Between the single vectors, we invoke the skip-gram approach to maintain the chronological relations among the tweets as illustrated in Fig. 4. The second input vector includes V_{n+1} to V_{n+m+1} tweet vectors (V_n) along with the first single vector includes V_n to V_{n+m} tweet vectors. Every k single input vectors are deemed as a batch. The LSTM model itself is the basic LSTM neural network based on TensorFlow. Here, we evaluate the dataset through the LSTM models of two scales. The basic parameters of the models are listed in Table III.

2) *Predicting Geo-Information and Tweet Content:* The softmax layer is utilized to predict the events according to the descending order of the probability of the words. The prediction only contains the words in the dictionary. To predict the geographic information that in which of the 9 areas the tweet is posted, we sum up the probability of each geographic elements in the single tweet vectors. As the output of the softmax layer is a $(k \cdot m \cdot (n+2)) \times vocabulary\_size$ matrix and the k, m, n are illustrated in Fig. 4. The vocabulary size is set to be 92000 to include all the indexes including both dictionary indexes and geographic indexes. Then for every $n+2$ rows (this represents a single tweet), we separately sum up the 81112^{th}, 81113^{th} and 81114^{th} columns of these $n+2$ rows. The greatest value among the three indicates that the tweet belongs to the corresponding latitude area. The determination of the longitude is of the same approach.

IV. PRACTICAL EXPERIMENTS FOR LATENT EVENTS FORECASTING

In this section, practical experiments are demonstrated to evaluate the three models utilizing the collected data from Twitter. The three models generate different types of outcomes: 1) The LDA model predicts latent events. 2) The LSTM model with skip-gram embedding forecasts related words of events. 3) The LSTM model with skip-gram-geo-aware embedding predicts related words of the event and its location.

A. Dataset Preparation

The Twitter dataset is collected between 27th January 2018 and 27th February 2018 (31 days in total) which composes approximately 70000 tweets with geo-tags in total. Compared with the parallel collecting experiment, which exploits the keyword "UK" to filter tweets and obtained 2 million tweets in total, the portion of tweets with geo-tags is approximately 2.89% of the tweets with keyword "UK". The coordinates of the tweets are restricted by the parameters as demonstrated above. The time range within a day of the collection is constrained from 7:00 AM to 15:30 PM (GMT).

1) *Distribution of the Data:* Due to the different distribution of population and BS in London, the magnitude of the tweets within different districts varies during the same time period as illustrated in the density map. We divide the entire London area equally into nine smaller areas by latitude and longitude of BS locations.

2) *Data Cleaning:* To satisfy the prerequisites of training the propose ML models with the dataset, some characters are deleted based on the standards below:

- Characters that do not belong to English.
- Punctuation and Stopwords (in the nltk.corpus package of Python).
- Numbers.

Since URLs in the tweets are largely points to specific objects, such as a web page or a piece of video, they are valuable in selecting caching contents. The URLs directing to the media files in tweets are stored separately with the tweet text. Therefore, it offers us an opportunity to associate the caching media files with their tweet text.

Particularly, the twitter tags, such as "#London", are not removed as they are able to be deemed as constrains of latent events. The reason that we employ tweet text and the tags rather than merely the tags is because 1) Not all tweets contains tags, some users are not accustomed to use tags. 2) The description of the files is not sufficient with tags—some tags are too vague to describe the media files. For instance, for a particular episode of a TV series, the tag is the whole TV series.

3) *Caching-Contents Datasets:* Regarding the fact that when a tweet contains certain text-related media files—images, videos, the URLs directing to the files are given when retrieving the tweets from Twitter API. Since we need to determine the exact caching contents based on tweet text, it is necessary to obtain the caching contents for the text datasets mentioned above. Among the approximately 70000 tweets with geo-tags in total, 7.69% of them contain media files(images and videos). As the tweets with videos are approximately 10 times more compared to the ones with videos, images take up a larger proportion in tweets' media files. Furthermore, since the videos in tweets are mainly short videos, the total occupied caching space of images is twice as large as that of videos. According to the reasons above, in this paper, we focus on caching both the images and videos from the tweets rather than merely the videos.

To demonstrate the performance of the proposed models, we adopt the perplexity value as the key performance indicator, because perplexity is an well accepted approach to evaluate natural language processing (NLP) models. The definition of perplexity is illustrated in Eq. (4):

$$2^{H(p,q)} = 2^{-\sum_{i=1}^{N} p(x_i)\log_b q(x_i)}, \tag{4}$$

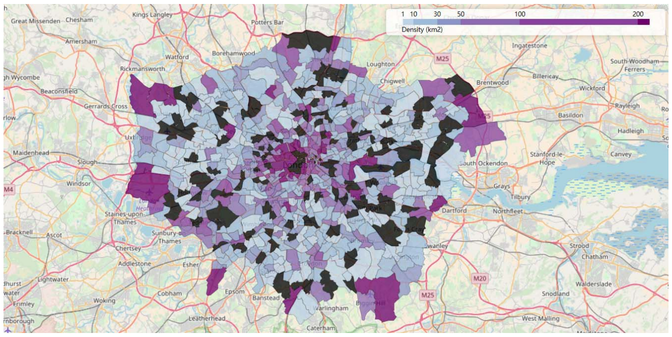


Fig. 5. The density distribution of the involved BSs in London (Area marked with deeper purple contains more BSs). We observe that the areas of City of London and Inner London are equipped with denser BSs compared with areas of Outer London.

where p is the unknown distribution of the test dataset, x_1, x_2, $x_3, \ldots, x_N$ are subsets of test dataset. q stands for the model that we want to evaluate. Perplexity evaluates the similarity between the prediction and the ground truth. Since perplexity maintains a reciprocal relationship with the Log-likelihood measures, a language model with lower perplexity achieves better performance during application. Therefore, in this section, we apply perplexity to demonstrate the accuracy of the proposed models and the testing datasets are clarified in each sub-section. Furthermore, the unified standard—complexity is also capable of generalizing the results since the standard is based on numerical results.

TABLE IV

SAMPLE RESULTS IN DIFFERENT REGIONS GENERATED FROM THE PROPOSED LDA MODEL

Region	***Topics***
1	park hyde loving fc tottenham gunners children
2	stadium teaching wembley fitness art centre 14
3	rain mm mph wind hpa thames fine
4	london greater heathrowairport station hounslow new city
5	job hiring england careerarc london latest work
6	park studio team join love old o2
7	uk egaylity 12 gay city stigmabase wembley
8	essex palace art teaching gallery north en
9	london greater bridge station house free unitedkingdom

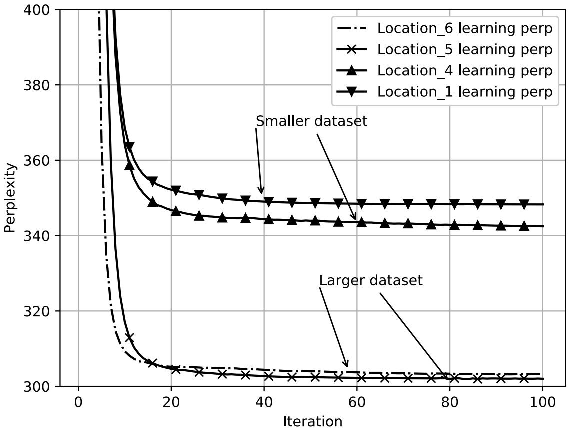


Fig. 6. The learning perplexity of the proposed LDA model in Location 1, Location 4 (smaller datasets), Location 5 and Location 6 (larger datasets).

B. LDA Model

1) Dataset Division: To demonstrate the distinct preference of topics within different regions, the dataset has been divided into 9 smaller datasets. The separation is based on the BS coordinates which is illustrated by Fig. 5. According to Fig. 5, areas of City of London and Inner London are equipped with denser BSs compared with areas of Outer London. The learning process of the proposed ML-base caching is performed in a centralized manner in the divided nine areas. We predict the latent events/topics for each of the divided nine area.

2) Literal Results: To demonstrate the variation of different extracted topics in different regions and considering the page limit, we display the results in Location 5 and Location 4 for examples. The reason for selecting the two locations is that the topics in the two regions are more diversified compared to other regions. The literal results of the LDA model in Location 5 are demonstrated below in Table IV. The results from the location 5 are able to be interpreted in the following approach. The first topic relates to employment advertising tweets with the tag of "#careerarc". The 2^{nd} and 3^{rd} topics associate with the equality of the "LGBT" group and the weather. According to the results in Table IV, the first and the third topic are possibly originated from the tweets composed by tourists as they are discussing the particular locations like the Heathrow Airport and Hounslow. The second topic is related to the Valentine's Day, however as there is also a music festival called "Kaleidoscope Festival" during the same time period, the keyword "kaleidoscope" is also included here.

The complexity tendencies of the LDA model under different datasets are illustrated in Fig.6. To illustrate the different tendencies of perplexity under datasets of different sizes, we select two larger datasets and two smaller datasets based on their regions. The datasets of location 6, location 5 are larger datasets while the rest two datasets are smaller ones. The graph demonstrate the results in two aspects. The decreasing trends demonstrate that the LDA model is able to converge and the larger datasets enable the model to achieve more accrate performance (lower complexity).

Here are we demonstrate some keywords of the events from LDA model as illustrated in Fig. 7. From the graph, tweets in the 9 areas are primarily related to the landmark places or district functions (sightseeing, sports), which assists to allocate different characteristics among the districts in London.

Analysis of the keywords is capable of determining the text-related caching contents based on the public preference. Since we have discovered that the keywords from tweets in different regions are associated closely with the indigenous landmarks, the text-related contents are able to be determined from them. For instance, according to the Fig. 7, in the south part, the keywords "wimbledon" and "stadium" are illustrated. Different from the conventional counting methods, our proposed ML approach is capable of demonstrate the concurrency among the words. Therefore, based on the two keywords mentioned above, the text-related contents, such as videos for

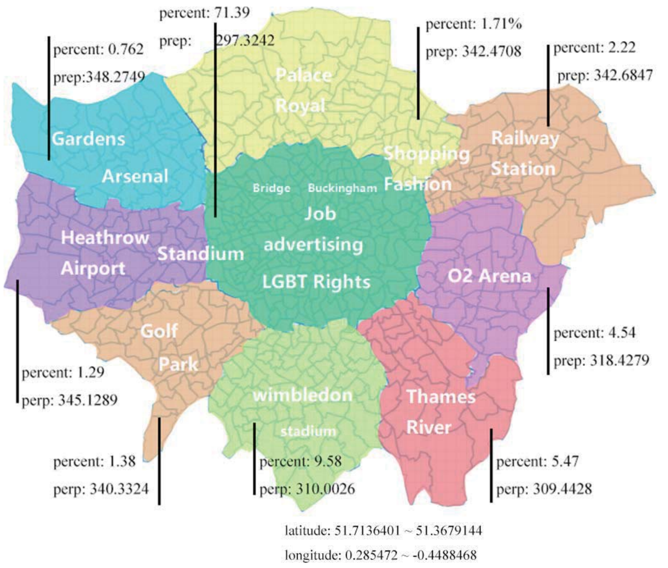


Fig. 7. Keywords distribution of the predicted events in London generated from the proposed LDA model.

the matches held in the stadium, are supposed to be cached at nearby base stations. Similarly, for the base stations located near the keyword "Arsenal", the related videos of the recent football games should be cached. Different from the situations mentioned above, regarding to the palaces in the central part of London, the pictures and navigation information (including coordinates and public transport information) should be cached since there are several sightseeing spots including the famous Buckingham Palace.

C. LSTM Model With Skip-Gram Embedding

1) Datasets: The 9 training datasets and 9 testing datasets are separated based on regions of BS. The training datasets range from 27^{th} January to 26^{th} February. The testing datasets are tweets on the 27^{th} February.

2) Forecasting Results: For the similar reason with the literal results in LDA model, we demonstrate the results from Location 4 and Location 5. The top events of the location 5 and location 4 are illustrated in the Table V. Same as LDA model, as the perplexity trends converges during the testing, the generalization and the accuracy of the model are able to be demonstrated. The interpretations are based on the keywords in the topics. With the keyword "job", "hiring", "#CEBCareers", the first topic is able to be considered relating to the job advertising while containing noise from irrelevant topics. The second topic involves a couple of landmarks, like "Belgravia", "Sleaford" and "Langham" along with meteorological terms like "UV", "Rain", "wind". Therefore the topic is considered discussing the weather corresponding to the landmarks. To generalize the results, numerical results are able to provide evidence. Since the perplexity of this model tends to converge and the final complexity values are satisfactory, the accuracy of LDA models are demonstrated.

While invoking the model on the location 5 dataset with networks of different scales, the variation of the training and testing perplexity is illustrated in Fig. 8.

TABLE V
SAMPLE RESULT EVENTS FROM THE PROPOSED LSTM MODEL WITH SKIP-GRAM EMBEDDING

Location	***Topics***
Location 5	#job #Hiring #CEBCareers May UK marriage Northern #BusinessMgmt urged Ireland take equal forward Manager #London anyone recommend Pourtsmouth #makeupartist May
Location 4	#London United Kingdom The Belgravia Click #ProductMgmt damn today mph fine UV Sleaford Langham England Rain We're #london #giftshop 2018

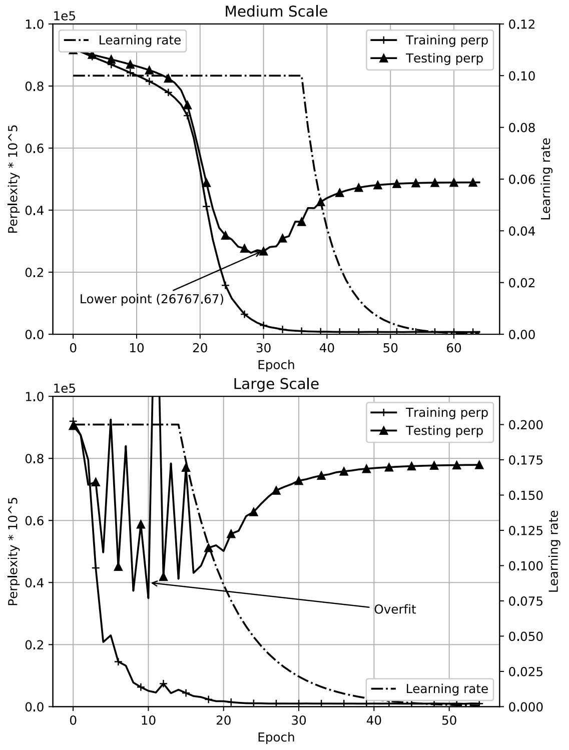


Fig. 8. The training and testing perplexity under medium scale and large scale of the proposed LSTM networks.

D. LSTM Model With Skip-Gram-Geo-Aware Embedding

1) Datasets: The training dataset contains all tweets from 27^{th} January to 26^{th} February. The testing dataset is the tweets on the 27^{th} February.

2) Forecasting Results: The training and testing perplexity tendency are illustrated in the Fig. 9. The text prediction examples are illustrated in Table VI. The topic is largely similar to the results of invoking the LDA model to the Location 5 dataset. As the job advertising tweets constantly contain a URL directing to their web site, the keywords "apply" and "click" are involved in the prediction. The prediction is more unified and of less noise compared to the results of the skip-gram embedding.

E. Model Analysis

In this section, the properties of the three models are analyzed. The performances of models under different circumstances are demonstrated. Based on the graphs mentioned

TABLE VI
RESULT TOPICS AND LOCATION EXAMPLES

Location Forecasting	***Topics***
Location 5	#London #CareerArc United #Hiring England The #job UK Kingdom work See latest Greater I'm opening Can apply Click

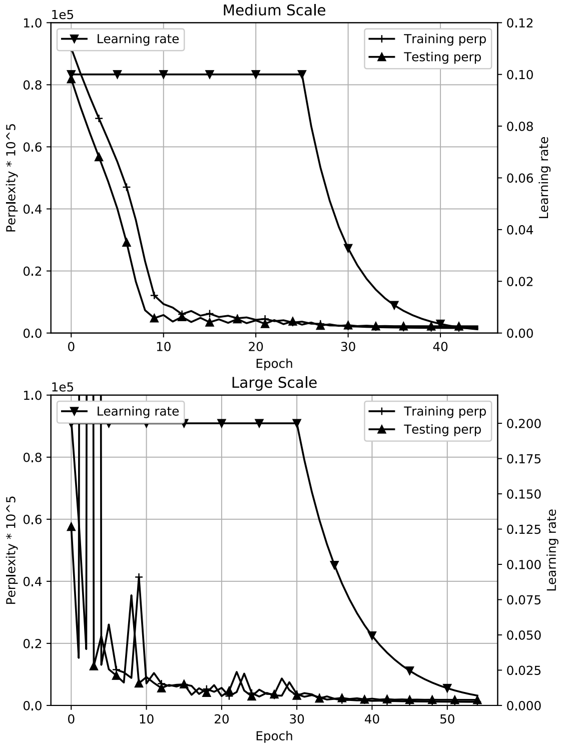


Fig. 9. The training and testing perplexity under medium scale and large scale of the proposed LSTM networks (Geo-aware).

in the above section, we discuss the different aspects of performance illustrated.

1) LDA Model With Different Datasets: The perplexity applied in the LDA model is able to be regarded as a standard to evaluate the cohesiveness between the extracted events and the ground truth—the existing tweet text. Since complexity is related tightly to the Log-likelihood, less complexity leads to more cohesiveness between the extracted events and the twitter text and more accuracy of the results. The convergence of the complexity is also able to be deemed the convergency of the model. With different scales of documents illustrated in Fig.6, larger documents (Location 5 / 6) tend to converge more rapidly. Furthermore, the overall complexity of larger documents tend to be lower. Since more tweets have involved and during each iteration, for definitive K latent events and a larger corpus, the event-word co-occurrence matrix is more reasonable. The final perplexity of the LDA model is satisfactory (about 300 under large dataset). Since the literal results demonstrated in Table IV contain correlative key words, we can interpret the events expediently. Regarding the advantages above, we demonstrate that the LDA model is feasible of predicting Twitter events.

2) Different Scales of LSTM With Skip-Gram Embedding: Different from the LDA model, less complexity of the model when applied on the testing datasets leads to less difference between the prediction and the ground truth—the testing datasets (future tweets). The convergence of the complexity is also able to deemed the convergency of the model. Under larger scale of network and learning rate, the tendency in Fig.8 tends to converge more rapidly while the testing perplexity fluctuates obviously due to overfitting and the characteristics of the gradient descend optimization. As mentioned above, complexity is applied to evaluate the accuracy of the models in this paper. According to the result, the smaller network and learning rate result in better overall complexity. The final testing complexity of the medium-scale network is 54738.23 while that of the large-scale network is 78312.39. Due to the polytropic combination of oral words, the perplexity of the Twitter documents are not restrained as that of the formal dataset like PTB (about 70 perplexity in large-scale LSTM model [37]).

3) Different Scales of LSTM With Skip-Gram-Geo-Aware Embedding: When the new model converges, the events of prediction are largely the same due to the characteristics of the LSTM model;therefore, the model itself is only capable of predicting one tweet in a specific area. The variations of the complexity under different scales of networks are demonstrated in Fig.9. The larger network with greater initial learning rate tends to converge more rapidly while fluctuating obviously. After the training and testing process, the larger network with greater depth of network structure concludes superior results on the overall testing perplexity. The outcome demonstrates that the larger network tends to generate more monolithic outcome results with the testing dataset. The final testing perplexity is 1253.75(Larger model) / 1752.83 (Medium model).

Based on the events prediction and perplexity results of the three models, the LDA model achieves satisfactory prediction as well as multiple events prediction within one particular area. While the basic LSTM model with skip-gram embedding is of relatively high perplexity. The novel LSTM model with skip-gram-Geo-aware embedding realizes comparatively lower perplexity and the geographic information prediction. However, due to the characteristics of the neural network when converges, future efforts are required to enable the model to predict multiple events.

V. ASSOCIATING EVENTS FORECASTING WITH CACHING

Three ML models are proposed in above sections to extract latent events from the existing tweets and predict future events based on existing tweets. The perplexity results confirm the effectiveness of the proposed models. Since we have proposed valid approaches to solve the core problems mentioned above, we need then to evaluate the model in the context of wireless caching. In this section, we associate the events prediction with the wireless caching. To exploit different advantages of different ML models, we propose the following approach to determine the text-related caching contents. While the LSTM-based models mentioned above are capable of predicting the future twitter text based on the chronological inputs of existing tweets, the LDA model is capable of extracting the latent events from the predicting twitter text. Afterwards, the text-related contents—the images and videos, which are capable of generating remarkable impact on the network backhaul loads, are cached at the BS. Since the accuracy of the framework is directly related to the accuracy of the LDA model, we mainly discuss the accuracy of LDA event-extracting model as the ML approach in this section.

The hit rates is applied to evaluate the performance of the proposed framework. Since determining text-related contents has no previous algorithms, we take conventional algorithms (LFU, LRU) for comparison. The LFU approach is applied to extract the most frequently used keywords from the existing tweet text while the LRU obtains the most recent keywords from the existing tweet text. The keywords extracted through the two conventional methods are applied to match actual media files. As for evaluation criteria, we proposed the "hit portion", which is the portion of utilized events among all the extracted events. Regarding the specific caching contents evaluation in this section, we demonstrate the difficulties of determining the caching contents of social media contents as well as propose our criterion to evaluate the model. Therefore, the coherency between the events and caching contents is better demonstrated. Here, the testing datasets are the tweets on 27th February while the training datasets are the tweets ranged from 27th January to 26th February.

A. TAC Strategy

In this section, we demonstrate the strategy for caching actual media files (images and videos) regarding the twitter topics.

1) Caching Objects: Since we have demonstrated the structure of our framework at the previous section and the approaches to obtain topics prediction have been demonstrated, we introduce the caching strategy to associate the topics with actual caching contents (images, videos). As the tweet text has been cleaned in the dataset preparation procedure, the words in tweet text are seldom meaningless and each of the words represents a symbol of identification. The core idea is to decide which tweet is valuable to be cached through predicted topics. To judge whether a topic is associated with a tweet, we define when there are 3 words the same between a topic and text of a tweet, the media files associated with this tweet are worthwhile to cache. After determining the tweet has the value to be cached—hit by the topics, the media files are cached through the URLs retrieved from the JavaScript Object Notation (JSON) data structure, which is obtained from Twitter API.

2) Prior List: After gathering the actual media files and extracting topics from existing tweet text, the relationship between the two parts is established. The popularity of the topics and actual caching objects is capable of being definitely settled through HTTP requests from users. Since there are generally large number of topics contained in tweet text and topics themselves are of popularity of different scale, the "Prior List" (PL) is applied to filter the topics of the most popularity to maximum the caching efficiency with certain number of topics and caching space for actual media files. Similarly, a PL of caching objects is a ranking list contains actual caching objects related to a specific topic obtained through Twitter API. The rankings of caching objects are based on their frequency of usage, namely their popularity. Regarding the existing caching strategy, which applies popularity of caching objects [21], the efficiency of the caching algorithm increases. With the aim of precisely determine what to cache, we create a PL for each topic to preserve the media files related to it. The objective of the PL is to associate the extracted topics with the popularity of the caching objects, which improves the efficiency of the caching framework.

As the BS is capable of monitoring the wireless traffic through HTTP requests / responses, PLs related to different topics are varying dynamically after deploying our caching framework at the BS. When the topics PL has reached maximum amount of topics and a new topic emerges, the popularity of the new topic is compared with those of topics in the topics PL. When the popularity of the new-emerged topic exceed any from those of the topics in PL, the PL is updated. Moreover, the caching space is updated.

B. Evaluating the Extracted Events

In this section, the performance of the ML approach as well as the conventional LFU, LRU approaches is evaluated through four different numerical results—tweet hit rate, tweet hit portion, cache portion and hit cache portion. The LFU and LRU approaches extract the keywords from the tweet text rather than the events like the ML approach. The aim of this section is to compare the performance of the models as well as evaluate the properties of the ML approach under different circumstances.

1) Tweet Hit Rate: In this part, we demonstrate hit rates of tweets in our model. The aim for considering tweet hit rates is to illustrate to what extent a model can associate the topics with the tweet text. The hit rate is calculated based on the number of hit topic and the number of all the topics, which is formally given by $P_{hit} = Numberofhittopicks/Numberofalltopics$. In Fig. 10, the hit rates of three approches are demonstrated—LFU, LRU and the ML approach. The X-axis of the graph is the amount of topics. By fixing the number of topics extracted from the tweet text, we are therefore capable of discussing the performances under different caching space since more topics lead to more media files to be cached. Rather than merely extracting keywords from the corpus like LFU and LRU, ML approach is actually extracting topics, i.e. the cohesive keywords from the corpus, which assist the approach achieve high coherency between the topics and the tweet text. Moreover, the proposed NOMA-ML approach is capable of finding more complex nonlinear relationship between collected twitter data and latent topics compared to conventional NOMA-LFU approach and NOMA-LRU approach. Therefore, the proposed NOMA-ML approach outperforms conventional NOMA-LFU approach and NOMA-LRU approach towards extracting topics of tweet text from the corpus.

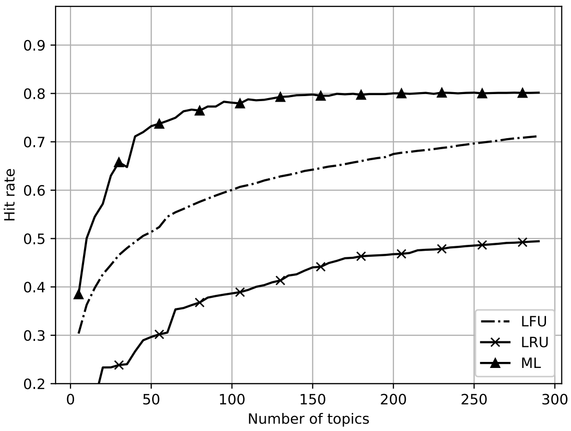


Fig. 10. The hit rate of topics of different approaches.

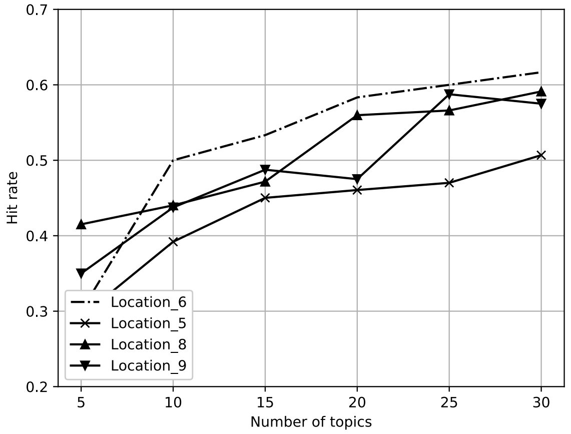


Fig. 11. The hit rate of topics in different regions.

With the aim of further developing the properties of the ML approach, we discuss the performance of the method under different circumstances. To demonstrate the different trends under datasets of different sizes, we select 4 datasets from different regions to illustrate the trends. The selection of the datasets are based on their sizes (number of tweets) with the aim to invest the model property in different dataset ranges. Among the 4 testing datasets exploited to evaluate the model, the sizes of training datasets are location 5 > location 8 > location 9 > location 6.

In Fig. 11, the trends of hit rates can be described as below. 1) The hit rates of the model on all testing datasets increases when the numbers of topics increases. 2) The overall hit rates vary little among the different sizesof datasets, which means that the little variance among the testing datasets does not influence the overall accuracy. 3) The highest hit rates are achieved when there are maximum number of topics and the highest hit rates vary from 50% to 65% under different testing datasets. As a conclusion, the wireless caching framework we propose is capable of generating relatively satisfactory results.

2) Tweet Hit Portion: To display the utilization ratio of the topics based on the three approaches mentioned above—LFU, LRU and ML approach. The tweet hit portion is the utilization ratio of the topics, namely the topics hit by the tweet text of the future. The aim of introducing this property is also to demonstrate the effectiveness and accuracy of the abilities to extract topical information from the corpus. We define the tweet hit portion to be the portion of topics hit by the testing tweet text dataset. The aim of this criterion is to illustrate the accuracy of the topics generated from different approaches. Besides, the performance of the ML approach under different datasets is discussed.

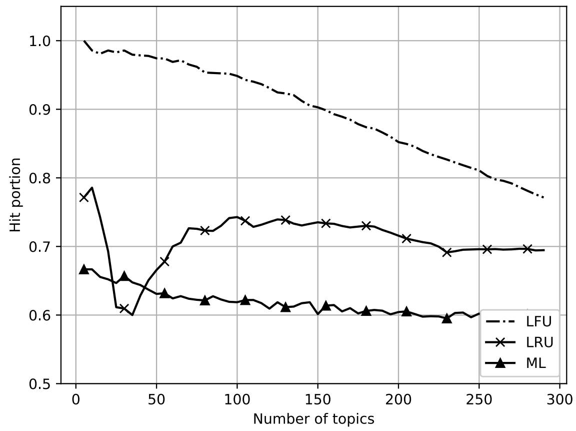


Fig. 12. The hit portion of topics of different approaches.

In Fig.12, the ML approach achieves the lowest hit portion rate while the LFU and LRU achieve higher hit portion. Therefore, the contradiction is that the ML approach achieves higher tweet hit rates as well as the lower hit portion under the same topics numbers compared to the other two methods. The explanation to this contradiction is that while the LFU and LRU focus on more monotonic topics—The recent ones and the hot ones, the topics from these two approaches are more unified compared to the ML approach. While the ML approach generates a more heterogeneous topics prediction, the hit portion is lower compared to LFU and LRU.

To invest the ML approach properties under datasets of different sizes, we choose the 4 datasets from different regions. The selection of the datasets are based on the familiar reason as that of the tweet hit rates section. In the Fig. 13, the trends of hit portions are demonstrated. The results are able to be explained from the following aspects. 1) The utilization ratios of the caching topics decrease along with the increasing number of extracted topics. 2) The over all trends of utilization ratios vary remarkably among different sizes of testing datasets. Moreover, larger datasets (location 5, 8 datasets) are able to achieve higher utilization ratios. 3) The highest utilization ratios are achieved when there are least extracted topics and the maximum utilization ratio is approximately 60% with the largest testing dataset of location 5.

C. Problem of Evaluation and Proposed Solution

During the collection of the datasets, we noticed an obstacle of evaluating the effectiveness and accuracy of our model. The tweets with geo-tags are largely the original ones, which are created by user themselves rather than retweet. This situation leads to the conclusion that the overlap of media files between

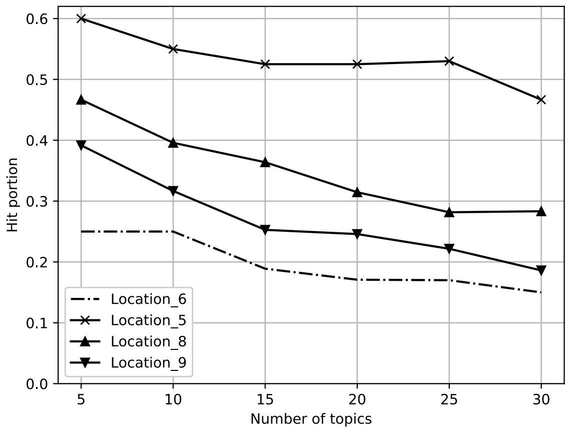


Fig. 13. The hit portion of topics in different regions.

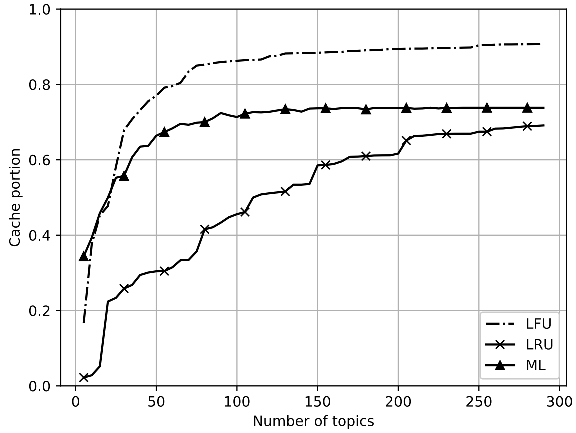


Fig. 14. The caching portion of the training dataset (different approaches).

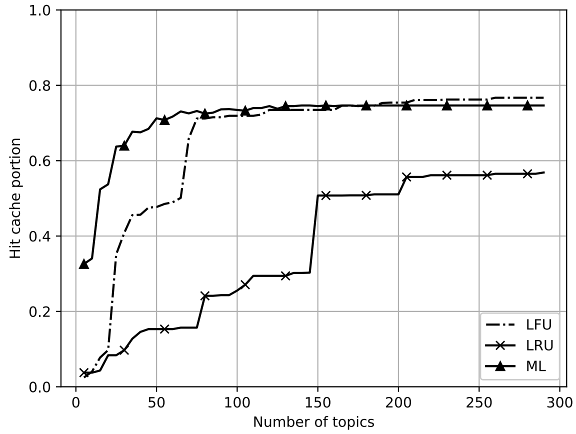


Fig. 15. The hit caching portion of the testing dataset (different approaches).

the two different days are particularly little. Different from conventional video sites, the contents of the social platforms are largely published by the public, which makes the determination of caching contents nontrivial. However, since the traffic of social media platforms is capable of being formulated into two aspects—viewing (downloading) and posting (uploading), we propose a unified approach to evaluate the model. Users prefer to not only go through the contents (text, images, videos) that they favor, but also to post their own contents associated to the topic, such as a video related to a popular kind of pet in that region. Therefore, when media contents of the next day are similar to the media contents at the present day, the media contents cached today are highly possible to be viewed by the users. Regarding to the reasoning above, we propose the two evaluation criteria, namely "cache portion" and "hit cache portion" to evaluate the models.

In this section, the performance of the models is evaluated through solid text-related caching contents(images, videos). To process the evaluation, the numerical results are presented from two different aspects.

1) Cache Portion: First part is the cache portion. Cache portion is the portion of media files (text-related caching contents) which are cached after the caching-content determination procedure. This criterion represents how much of the formal caching contents is cached based on the topics, which relates tightly to the size of occupied caching space at BS. With higher caching portion, the requirements for caching space are higher—more contents are cached. Here, the portion is the size of files that is cached divided by the total size of media files in the training dataset.

In Fig.14, the cache portion of the 3 approaches—LFU, LRU and ML approach are illustrated. In the figure, LFU maintains the highest cache portion up to about 90%, which leads to the conclusion that under our caching strategy, 90% of the existing contents are cached to fulfill the future needs. While LRU achieves the least cache portion (70%), ML approach maintains the medium caching portion as approximately 75% of the existing media files. The curve of the ML approach also demonstrates that when the number of topics is restricted to 100-150, the ML approach achieves the stable status—increasing of the topics number results in little increase of the contents to be cached. This part results cooperate with the "hit cache portion" to obtain further conclusions.

2) Hit Cache Portion: In this section, the hit cache portion is employed as the criterion to evaluate the models. Hit cache portion is portion of how much footprint of media files in the testing dataset is hit by the obtained topics. As we mentioned before, since conventional "hit rate" measurement is not suitable for Twitter caching scenario, we propose this criterion to demonstrate the coherency between the predicted events and the future media contents in the testing dataset.

In Fig.15, the hit cache portion of the three approaches has been demonstrated. Regarding the graph, LFU and ML achieves much better performance (approximately 75%) compared to the LRU (less than 60%). This leads to the conclusion that the ML approach and the LFU approach are feasible of associating the obtained events with the actual caching contents. However, considering the results from the "cache portion" section, LFU approach is actually costing more footprint to achieve the satisfactory result. This leads to the conclusion that our ML approach caches less redundant contents compared to the conventional LFU method. The reason for not achieving 100% caching accuracy is that some caching contents in the testing dataset are irrelevant to the existing media files, while leads to the situation that the events are not capable of being associated with this part of contents.

To reach a conclusion, the ML approach we proposed is evaluated through several different criteria against two conventional caching algorithms. With higher tweet hit rate, our proposed ML approach is capable of achieving satisfying events prediction results from the tweet text aspect. Regarding to the caching contents evaluation we demonstrate above, the ML approach is feasible of achieving high consistency between the events and the future caching contents. The cache portion results illustrate that the ML approach achieves less caching redundancy.

D. Content Delivery

For transmit the contents for multiple users simultaneously, we conceive a non-orthogonal multiple access (NOMA) downlink strategy for content transmission. Compared with conventional orthogonal multiple access (OMA) based content transmission, NOMA based content transmission for the proposed framework has high bandwidth efficiency and ultrahigh connectivity [39], when there are massive users requesting for content delivery. We consider a two-user NOMA downlink scenario. Without loss of generality, we assume that the channel gain of users satisfies $|h_n|^2 > |h_m|^2$. Assuming that the users request the same content from the BS, based on Shannon's capacity formula, the achievable rate of the users are given by

$$\mathrm{R}_n^{\mathrm{NOMA}} = B\log_2\left(1 + \frac{\rho_n |h_n|^2 {r_n}^{-\alpha}}{BN_0}\right), \tag{5}$$

and

$$\mathrm{R}_m^{\mathrm{NOMA}} = B\log_2\left(1 + \frac{\rho_m |h_m|^2 {r_m}^{-\alpha}}{\rho_n |h_m|^2 {r_m}^{-\alpha} + BN_0}\right), \tag{6}$$

where B is the bandwidth. ρ_m and ρ_n are the transmit powers of the users. ${r_m}^{-\alpha}$ and ${r_n}^{-\alpha}$ represent standard distance-dependent power law pathloss attenuation between the BS and the users. N_0 is the noise spectral density. We assume that all the videos associated with the twitter event have same size, dented by S. Then, the transmission delay is given by $T_1 = \rho_{ca}\frac{S}{\mathrm{R}_1}$, where ρ_{ca} is the hit rate of a caching scheme and R_1 is the transmission data rate of the user, which is defined in equation (5) and (6).

Fig. 16 presents the transmission delay of the proposed ML and the benchmarks. Fig. 16 is obtained assuming a downlink transmission scenario, where one BS serves two mobile users. We assume the bandwidth is 20MHz and all the associated videos have the same size, which is 10M. The distances between the users and the BS is 100m and 500m. According to Fig. 10, when the number of topics is 100, the hit rates are 0.8, 0.6 and 0.4 for ML, LFU and LRU. According to Fig. 16, the proposed NOMA-ML scheme outperforms conventional OMA, LFU and LRU schemes. It is also noted that the proposed NOMA-ML outperms NOMA-LFU scheme by 25%. According to Fig. 16, NOMA enabled transmission delay is also lower than conventional OMA enabled transmission delay when the number of users increases, which means that the performance of the proposed framework also has satisfactory performance when the number of users is more than two.

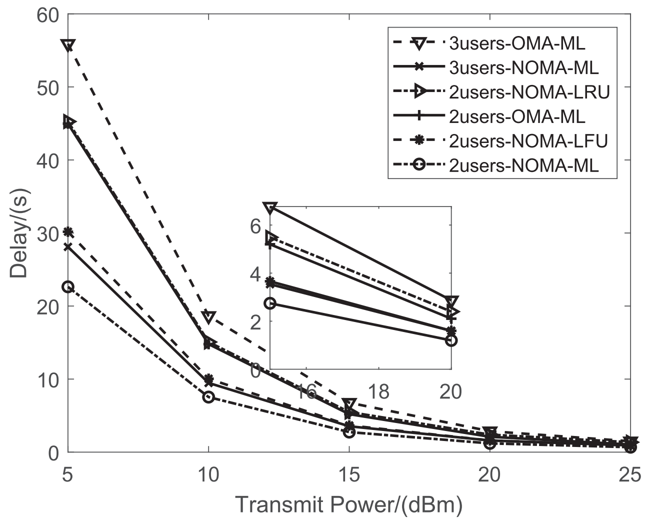


Fig. 16. The transmission delay under different caching schemes.

VI. Conclusion and Future Work

In this paper, a novel Twitter aided content caching (TAC) framework was proposed, which associated the tweets with the BS information. To associate Twitter events with the relative BS, the dataset was established to map tweets to their corresponding BS. Three machine learning (ML) approaches of allocating Twitter events with geographic information of BS were evaluated. Compared to LSTM model with skip-gram embedding, LDA and LDA-based approaches were capable of generating satisfactory predicting results in different regions. Regarding the results and the tendency of perplexity, our novel LSTM model with skip-gram-Geo-aware embedding was compatible to process the tweets with BS information. With the aid of ML based wireless caching techniques, the redundancy of the caching content was diminished. The effectiveness of the proposed solution was illustrated by practical experiments. However, since the characteristics vary among different dataset. One promising extension of this work is to investigate the performance of the proposed caching schemes to other social media datasets and open-accessed datasets.

References

[1] Y. Qi, Z. Yang, Z. Qin, Y. Liu, and Y. Chen, "Big data prediction in location-aware wireless caching: A machine learning approach," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[2] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022," Cisco, San Jose, CA, USA, White Paper, Feb. 2019. [Online]. Available: https://s3.amazonaws.com/media.mediapost.com/uploads/CiscoForecast.pdf

[3] A. Ioannou and S. Weber, "A survey of caching policies and forwarding mechanisms in information-centric networking," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2847–2886, 4th Quart., 2016.

[4] M. Tang, L. Gao, and J. Huang, "Communication, computation, and caching resource sharing for the Internet of Things," *IEEE Commun. Mag.*, vol. 58, no. 4, pp. 75–80, Apr. 2020.

[5] C. Wang, C. Liang, F. R. Yu, Q. Chen, and L. Tang, "Computation offloading and resource allocation in wireless cellular networks with mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 4924–4938, Aug. 2017.

[6] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. 18th Annu. Joint Conf. IEEE Comput. Commun. Soc. (INFOCOM)*, vol. 1. Mar. 1999, pp. 126–134.

[7] Statista, "Number of social network users worldwide from 2010 to 2023," Statista, Hamburg, Germany, White Paper, Jul. 2019. [Online]. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users

[8] K. C. Tsai, L. Wang, and Z. Han, "Caching for mobile social networks with deep learning: Twitter analysis for 2016 U.S. Election," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 1, pp. 193–204, Jan. 2020.

[9] A. Xiao, X. Huang, S. Wu, C. Jiang, L. Ma, and Z. Han, "User preference aware resource management for wireless communication networks," *IEEE Netw.*, vol. 34, no. 3, pp. 78–85, May 2020.

[10] K. Zhu, W. Zhi, X. Chen, and L. Zhang, "Socially motivated data caching in ultra-dense small cell networks," *IEEE Netw.*, vol. 31, no. 4, pp. 42–48, Jul. 2017.

[11] E. Zeydan *et al.*, "Big data caching for networking: Moving from cloud to edge," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 36–42, Sep. 2016.

[12] G. Lansley and P. A. Longley, "The geography of Twitter topics in London," *Comput., Environ. Urban Syst.*, vol. 58, pp. 85–96, Jul. 2016.

[13] B. O'Connor *et al.*, "From tweets to polls: Linking text sentiment to public opinion time series," in *Proc. AAAI Conf. Weblogs Social Media (ICWSM)*, May 2010, vol. 11, nos. 122–129, pp. 1–2.

[14] S. T. Piantadosi, "Zipf's word frequency law in natural language: A critical review and future directions," *Psychonomic Bull. Rev.*, vol. 21, no. 5, pp. 1112–1130, Oct. 2014.

[15] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1107–1115.

[16] K. Zhang, S. Leng, Y. He, S. Maharjan, and Y. Zhang, "Cooperative content caching in 5G networks with mobile edge computing," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 80–87, Jun. 2018.

[17] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[18] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[19] H. Pang, J. Liu, X. Fan, and L. Sun, "Toward smart and cooperative edge caching for 5G networks: A deep learning based approach," in *Proc. IEEE/ACM IWQoS*, Jun. 2018, pp. 1–6.

[20] E. Baştuğ, M. Bennis, and M. Debbah, "A transfer learning approach for cache-enabled wireless networks," in *Proc. 13th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, May 2015, pp. 161–166.

[21] Z. Yang, Y. Liu, Y. Chen, and L. Jiao, "Learning automata based Q-Learning for content placement in cooperative caching," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3667–3680, Jun. 2020.

[22] X. Zhao, P. Yuan, H. Li, and S. Tang, "Collaborative edge caching in context-aware device-to-device networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9583–9596, Oct. 2018.

[23] S. Müller, O. Atan, M. van der Schaar, and A. Klein, "Context-aware proactive content caching with service differentiation in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 2, pp. 1024–1036, Feb. 2017.

[24] D. Zheng, Y. Chen, M. Yin, and B. Jiao, "Cooperative cache-aware recommendation system for multiple internet content providers," *IEEE Wireless Commun. Lett.*, vol. 9, no. 12, pp. 2112–2115, Dec. 2020.

[25] Y. Fu, Q. Yu, T. Q. S. Quek, and W. Wen, "Revenue maximization for content-oriented wireless caching networks (CWCNs) with repair and recommendation considerations," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 284–298, Jan. 2021.

[26] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

[27] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 938–953, Apr. 2016.

[28] B. Yang, W. Guo, B. Chen, G. Yang, and J. Zhang, "Estimating mobile traffic demand using Twitter," *IEEE Wireless Commun. Lett.*, vol. 5, no. 4, pp. 380–383, Aug. 2016.

[29] D. Ramage, S. T. Dumais, and D. J. Liebling, "Characterizing microblogs with topic models," in *Proc. AAAI Conf. Weblogs Social Media (ICWSM)*, May 2010, vol. 10, no. 1, p. 16.

[30] J. Wang and M. She, "Probabilistic latent semantic analysis for multichannel biomedical signal clustering," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1821–1824, Dec. 2016.

[31] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook Latent Semantic Anal.*, vol. 427, no. 7, pp. 424–440, 2007.

[32] A. Fang, C. Macdonald, I. Ounis, and P. Habel, *Topics in Tweets: A User Study of Topic Coherence Metrics for Twitter Data*. Cambridge, MA, USA: MIT Press, 2016.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[34] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 3104–3112.

[35] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

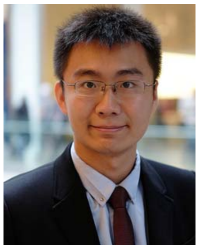
[36] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1724–1734.

[37] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," 2014, *arXiv:1409.2329*. [Online]. Available: http://arxiv.org/abs/1409.2329

[38] N. M. Nasrabadi, "Pattern recognition and machine learning," *J. Electron. Imag.*, vol. 16, no. 4, 2007, Art. no. 049901.

[39] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and L. Hanzo, "Enhancing the physical layer security of non-orthogonal multiple access in large-scale networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1656–1672, Mar. 2017.

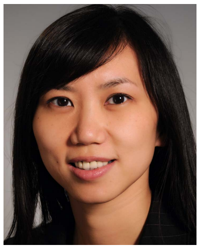
Zhong Yang (Student Member, IEEE) received the B.S. degree from the Nanjing University of Science and Technology (NUST), Nanjing, China, in 2013. He is currently pursuing the Ph.D. degree with the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests include machine learning, non-orthogonal multiple access, reconfigurable intelligent surface, mobile edge computing, and caching.



Yuanwei Liu (Senior Member, IEEE) received the B.S. and M.S. degrees from the Beijing University of Posts and Telecommunications, in 2011 and 2014, respectively, and the Ph.D. degree in electrical engineering from the Queen Mary University of London, U.K., in 2016.

He was with the Department of Informatics, King's College London, from 2016 to 2017, where he was a Post-Doctoral Research Fellow. He has been a Lecturer (Assistant Professor) with the School of Electronic Engineering and Computer Science, Queen Mary University of London, since 2017. His research interests include non-orthogonal multiple access, 5G/6G networks, machine learning, and stochastic geometry. He has served as a TPC Member for many IEEE conferences, such as GLOBECOM and ICC. He received the IEEE ComSoc Outstanding Young Researcher Award for EMEA in 2020 and the 2020 Early Achievement Award of the IEEE ComSoc Signal Processing and Computing for Communications (SPCC) Technical Committee. He also received the Exemplary Reviewer Certificate of IEEE WIRELESS COMMUNICATIONS LETTERS in 2015, IEEE TRANSACTIONS ON COMMUNICATIONS in 2016 and 2017, and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS in 2017 and 2018. He is currently a Senior Editor of IEEE COMMUNICATIONS LETTERS, and an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and the IEEE TRANSACTIONS ON COMMUNICATIONS. He serves as the Leading Guest Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (JSAC) Special Issue on Next Generation Multiple Access, and a Guest Editor for IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING (JSTSP) Special Issue on Signal

Processing Advances for Non-Orthogonal Multiple Access in Next Generation Wireless Networks. He has served as the Publicity Co-Chair for VTC 2019-Fall. He is the leading contributor for best readings of non-orthogonal multiple access (NOMA) and the primary contributor for best readings of reconfigurable intelligent surfaces (RIS). He serves as the Chair for the Special Interest Group (SIG), SPCC Technical Committee on the topic of signal processing techniques for next generation multiple access (NGMA), the Vice-Chair of the SIG, Wireless Communications Technical Committee (WTC) on the topic of reconfigurable intelligent surfaces for smart radio environments (RISE), and the Tutorials and Invited Presentations Officer for Reconfigurable Intelligent Surfaces Emerging Technology Initiative.



Yue Chen (Senior Member, IEEE) received the bachelor's and master's degrees from the Beijing University of Posts and Telecommunications, Beijing, China, in 1997 and 2000, respectively, and the Ph.D. degree from the Queen Mary University of London (QMUL), London, U.K., in 2003. She is currently a Professor of telecommunications engineering with the School of Electronic Engineering and Computer Science, QMUL. Her current research interests include intelligent radio resource management for wireless networks, cognitive and cooperative wireless networking, mobile edge computing, HetNets, smart energy systems, and the Internet of Things. She has served as a TPC Member for many IEEE conferences, such as GLOBECOM and ICC. She is currently serving on the Editorial Board as an Editor of the IEEE COMMUNICATION LETTERS.

Joey Tianyi Zhou (Senior Member, IEEE) received the Ph.D. degree in computer science from Nanyang Technological University (NTU), Singapore. He is currently a Senior Scientist, a Principal Investigator, and a Group Manager with the Institute of High Performance Computing (IHPC), Agency for Science, Technology, and Research (A*STAR), Singapore. He is also leading the AI Group with more than 30 research staff members. He is holding an Adjunct Faculty position with the National University of Singapore (NUS). Before working at IHPC, he was a Senior Research Engineer with the SONY US Research Center, San Jose, CA, USA. His current interests mainly focus on machine learning with limited resources and their applications to natural language processing and computer vision tasks. He received the NIPS Best Reviewer Award in 2017. He organized ICDCS'20-21 workshop on efficient AI meets edge computing, ACML'16 workshop on learning on big data workshop, and IJCAI'19 workshop on multi-output learning. He has served as an Associate/Guest Editor for IEEE ACCESS, *IET Image Processing*, IEEE MULTIMEDIA, and *ACM Transactions on Multimedia Computing, Communications, and Applications* (TOMM), *Springer Nature Computer Science* and the TPC Chair in Mobimedia 2020.