ISSN 2976-1190 (Online)
Journal of Engineering and Sciences
Vol. 1 Issue 1

Automatic Identification of Nepalese Tourism Related Tweets using Machine Learning Algorithms

Bhawana Poudel[1] , Bidur Devkota[2] , Nabin Lamichhane[2]

[1] *Department of IT, Gandaki University, Nepal*
[2]*Department of Electronics and Computer Engineering, IOE, Pashchimanchal Campus, Nepal*

(Manuscript Received 16/08/2022; Review: 30/09/2022; Revised 03/10/2022; Accepted 05/10/2022)

Abstract

With the widespread use of social media, large volume of user generated public data is readily available for research use. These media disseminate spatiotemporal public opinion regarding range of events, activities and human behaviors. Such data can be mined to get better insights in many fields like tourism, health, marketing, etc. Nepal being a growing hub of tourism, such data can be explored to uncover valuable insights related to tourism. This study focuses on examining such public views expressed in the microblogging and social networking site, Twitter. The focus is to illustrates a way to bridge that gap via automated identification of Nepal tourism-related tweets with the help of supervised machine learning methods i.e. Support Vector Machines and Naïve Bayes. Such algorithms determine a target category of a tweet by examining the tweet words and the desired target categories. In this way, supervised classification methods automatically assign a particular tweet to a predefined category, i.e. tourism-related or not. As a first step in this work, tweet dataset is developed from tweet corpus by collecting and labeling tweets as Nepal tourism related or not. The proposed approaches yielded promising results in terms of different performance metrics such as precision, recall, F1 score and accuracy. Based on the classification report, Support Vector Machines outperformed Naïve Bayes algorithm. Standard performance measures showed an overall accuracy of 91% and 81% and F-scores of 0.91 and 0.77 respectively for Support Vector Machines and Naïve Bayes. The outcomes of this study can provide valuable insights regarding the Nepalese Tourism Brand presence in social media and can act as an starting point for further in depth investigations. It can provide a valuable reference for various stakeholders such as tourism planners, urban planners, and so on.

Keywords: Naive Bayes; Nepal Tourism; SVM; Text Classification; Tweets.

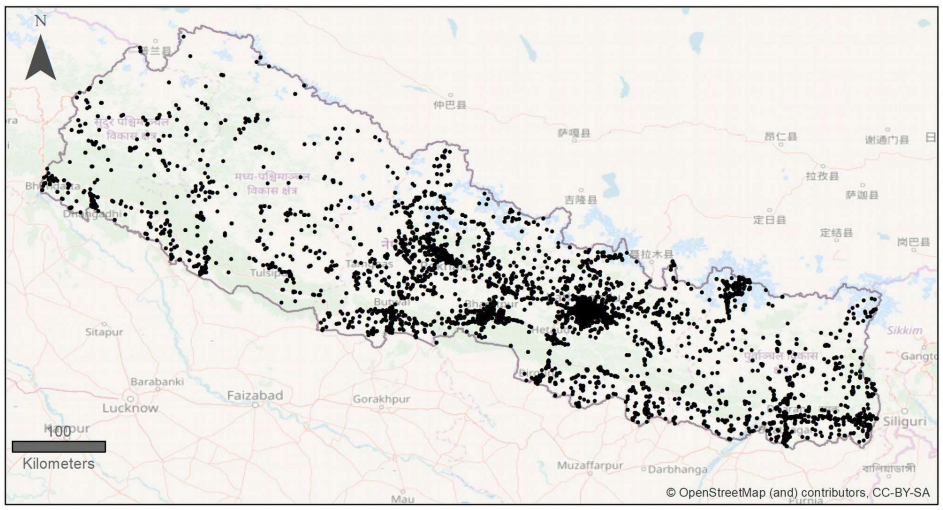

Figure 1 : Distribution of geo-tagged tweets in Nepal [24]

1. Introduction

The advent of Online Social Networks (OSNs) such as Twitter, Facebook and LinkedIn have

changed the entire way of human life; in the way we behave, we learn, we communicate, we interact and we live. Basically, Online Social Networks utilize Internet programs to connect and interact with various like-minded people and organizations. Eventually, online social media have evolved into a powerful communication tool. The services provided by them have opened near limitless possibilities in various areas such as maintaining online connection with friends and new acquaintances, create and share things like blogs, images, videos, comments and so on [1]. Recent research shows that, around 58% of the people have a profile for a social network. Facebook have the largest user base and following it were LinkedIn, Twitter and Google Plus. The use of these social networks generates a large volume of data in various formats [2].

In July 2006, Jack Dorsey launched a OSN site called Twitter [3]. It provides a micro blogging service that facilitates users to publish a text message up to 280 characters which is popularly known as tweet. It is one of the top ten most visited sites in the Internet [4]. It is a service that a large number of people from various sectors use to disseminate real-time user generated content to the world. An average of 58 million tweets containing rich data like texts, images, links and videos are posted every day in Twitter [5]. Generally, such data not only deliver various kinds of information that the user explicitly writes but also provides valuable information relating to the user profile and location. Such a wealth of information provides a great working ground for querying and mining noble insights and developing innovating applications. Some of the notable examples include personalized news recommender system based on user profile [6] , extracting major life events from twitter data [7], real time event detection [8] and so on.

The advent of the new way of information sharing via online social media has helped the tourism industry in a great way. Volunteered Geographic Information sites such as Twitter and TripAdvisor are being used as an easy platform for the market promotion and live interactions among the travelers and the related service providers. Moreover, they are playing a significant role in changing the way travelers search, discover, share and rely on the information in order to plan and tailor the travel routes and destinations. The user generated contents sharing travelers' experiences acts as a word-of-mouth for information seekers which motivate them more than the information provided by the marketers and service providers via conventional information channels. A finding of a research by PhoCusWright suggests that, 90% of the cyber travelers rely on online reviews related to the tourism services and products [10]. Lately, the penetration of GPS receiver enabled mobile devices has increased. In 2009, Tweet service was enriched with the facility by which users could reveal the location of their tweets [11]. This has availed application developers and researchers with a wealth of location specific tweets called geo-tagged tweets. Geo-tagging tweets can be done in two ways. One way is to manually provide the location by specifying the place name like city. Another approach comparatively more precise than the first approach is to use the GPS device to extract and embed the location coordinates in the tweets [12]. Geo-tagged tweets are the twitter user posts which contains the corresponding geographical identification metadata like latitude and longitude. Moreover, geo-tagged tweets include a huge amount of data referring to time, place, event and so on. Hence, they are regarded as useful sources for eliciting tourism related information. Though there are many services provided by social networking sites, we are not being able to utilize them properly in proper industry.

Nepal is considered as one of the best international tourist destinations in the world. Also, it is one of the major sources of Nepalese income. However, the tourism business in Nepal is not explored enough in the international market. Most of highly potential but unexplored tourist destinations are not advertised and promoted effectively so that large number of probable tourists does not know about Nepal in international front. Though Nepal has accomplished various activities for the promotion of Tourism various domestic and international market by celebrating “visit Nepal 1998”, “Nepal Tourism Year 2011”, “Visit Lumbini Year 2012”, and “Everest Diamond jubilee 2013” [13] [14], much of those activities have not utilized the services of social networking sites like Twitter in

an innovative way. The social networking related researches in the tourism industry are still in its infancy. It is difficult to investigate on the impact of social media on all aspects of the tourism industry including local communities, and to demonstrate the economic contribution of social media to the industry. According to Nepal Tourism Vision 2020, Nepalese government intends to develop tourism to 2 million by 2020. So, one of the objectives of Nepal Tourism Vision is branding the Nepalese Tourism in national and international market via various promotional activities. One of the targeted media for promotion is Internet marketing. However, tourism growth is dependent on a number of factors such as development and improvement of infrastructure, information, facilities, access, transportation options, safety and security which are all needed in the case of Nepal [9].

In this research, Nepal tourism related contents are extracted and analyzed. This study concentrates two kinds of problems i.e. creating a dataset of tweets and devising a way to investigate the Nepalese tourism brand presence in social media by classifying tweets as Nepalese tourism related or not. It seems interesting to identify if the text posted in Twitter is related to Nepalese Tourism or not. This will provide some idea about the branding and presence of Nepal Tourism in Twitter. Various machine learning algorithm are available for identifying whether the tweets are related to Nepalese Tourism or not which is explored in this study. Though the literature avails various studies for tweet categorization, this work is significant as we develop a dataset with Nepalese tourism related tweets and use machine learning algorithms, i.e. Naive Bayes and SVM, to automate the identification of Nepal tourism related social media contents. Further, classification performance of these algorithms was also done.

2. Related Works

2.1 Twitter Data in Research Use

The travel and tourism industry have been utilizing the services of Online Social Networks for various activities. OSNs have proved to be an excellent means as an electronic word-of-mouth for spreading recommendations and opinions. Travelers rely on search engines and online social platforms in planning and customizing their travel itineraries. A number of studies have been accomplished to provide insights on how OSNs like twitter can assist the travel and tourism industry. Next, we present an overview of some literature illustrating the use of social media in Nepal and worldwide.

A study titled “Online Social Networks and its potentialities for Nepalese Tourism Promotion” recommends that Nepalese Tourism sector should consider deeper on using OSNs like Facebook and Twitter, as an effective and cost-efficient marketing and promotional tool [19]. The related agencies are encouraged not only to create mere social media presence but also expand their promotional activities by utilizing the hidden information available in the user generated contents. In an investigation [22], significant tweet clusters are observed in the settlement areas, major urban centers and natural areas. An interesting fact is that a major portion (i.e. one-third) of the tweets in Nepal are contributed by the foreigners. Further, an examination of the Twitter presence of the different country users and the foreigner arrival data was done and a positive correlation was discovered. This indicates a good relation between Twitter activities and tourism in Nepal. Therefore, it is reasonable to investigation ways for automatic identification of Nepal tourism related tweets. A study titled “Using Volunteered Geographic Information and Nighttime Light Remote Sensing Data to Identify Tourism Areas of Interest” published the distribution of geo-tagged tweets collected for 685 days via free streaming Twitter API [24]. This study examines spatial clusters of social media activities in different regions of Nepal identifies tourism areas of interest. Further exploration of the social media data was done in “Utilizing User Generated Contents to describe Tourism Areas of Interest” [23]. The study proposed a method to describe tourism area of interest (TAOI) by aggregating user generated social media text from Twitter and Flickr. The recommended bottom-up approach enables the

extraction of valuable information that is not possible by using traditional top-down approaches. Promising results were obtained while applying the proposed approach against popular tourism spots in the Kaski district of Nepal.

Alowibdi and the team performed an investigation and developed a location-based application called Vacation Finder. It can suggest potential vacation to the general users and also provide a guideline for the government body regarding tourism promotion in the country. Their technique relied on indexing, spatio temporal querying and machine learning. The overall process includes collection, analysis and visualization of geo-tagged tweets so as to help the users by suggesting a place for vacation [15]. A research entitled "Mapping Geotagged Tweets to Tourist Spots Considering Activity Region of Spot" was performed to develop a recommender system for tourist spots. Geotagged tweets were mined to characterize tourist spots. One-class support vector machine was used to detect the areas of significant activity in the proximity of target spots by using geotagged tweets and photographs. The unknown geotagged twitter posts from activity regions were taken as input which was then mapped to the target spots [16]. Shimada et. al. performed a study on analyzing tourism information on twitter in the context of a local city. Particularly, the study concentrated on fundamental technologies for the tourism information analysis system, i.e. tourism information extraction and sentiment analysis of the extracted information. The proposed approach extracted tweets related to a specific location and tourism events. Further, opinion mining of the tweet was performed using unsupervised machine learning approach based on a naive Bayes classifier [17]. Shimada et. al. further investigated on twitter and tourism in their study entitled "On-site Likelihood Identification of Tweets for Tourism Information Analysis". They analyzed tweets to extract tourism information. The extraction process is normally based on eliciting sentences that contains keywords related to desired facilities and events. Some of the sentences extracted in such way may not be relevant to tourism. Hence, this study proposed a machine learning technique for measuring the likelihood of tourism information [18].

2.2 Text Classification Algorithms

Classification groups input text into categories. It aims to determine a target category of a document by examining the words in the input document and the available target categories. In this way, text classification automatically assigns a specified document to one or more predefined category. Hence, it has become a promising approach to manage and organize a massive amount of documents.

Various researchers have demonstrated that text analysis method that facilitates text classification. Some of the popularly used algorithms for text classifications are Naive Bayes and Support Vector Machine. The Support Vector Machine is a classifier that finds the best hyperplane between two classes of data by separating positive an negative examples through the solid line in the middle called decision line [27]. Naive Bayes Classifier is basically a probabilistic classifier based on hypothesis. On the basis of assumption and training document, Bayesian learning is to find most appropriate assumption based on prior hypothesis and initial knowledge. Main assumption is that terms in test document have no relation among them and probability is calculated that document belong to category C [28]. Zubrinic and the research team performed a comparative study of Naive Bayes and SVM Classifiers in Categorization of Concept Maps [26]. The study proposed a way for automatic classification of concept maps using bag of words model. The study examined the ability of classification of concept maps using bag of words approach. The best results are achieved using multinomial Naive Bayes classifier.

The main focus of our proposed study is to explore and investigate with Nepal Tourism related tweets. Till date, there are no any previous study on Nepal Tourism related tweet classification. Hence, applying classification algorithms like Naive Bayes and SVM to such tweets is a noble work. This kind of study of tweets for investigating the brand presence of Nepalese Tourism has not be done to the best of our knowledge.

Figure 2: Tweet Classification Methodology

3. Methodology

Figure 2 shows the general method for tweet Classification Here, tweets are collected to develop Nepal tourism dataset. The dataset contains a collection of tweets related and not related to Nepal tourism. Tweet preprocessing is applied for removing noise and characters like smilies, stop words, etc. Preprocessing activities also include steps like tokenization and stemming of the tweet. Then, Naive Bayes and SVN algorithm were applied to the tweets and the classification report is examined.

3.1. Tweet Extraction and Preprocessing

The Nepal Tourism related Tweet dataset is not available to the best of our knowledge. Hence, developing the dataset is an important task. Popular Nepal related keywords are examined and such keywords are matched to collect the Nepal tourism related tweets. The preprocessing of the tweets ensures that the tweets are noise free and ready for applying the classification algorithms. Stop word filtering, special character filtering and language filtering are some of the important activities performed during the preprocessing. Only English language text are used in this study. NLTK's list of English stop words are used to remove common words that have no significance in the classification.

3.2. Tweet Classification

Supervised machine learning approach is used for tweet text classification in order to automatically categorize a tweet into predefined labels, i.e. "Related" or "notRelated" to tourism. The probability for a tweet to be in a particular category is calculated based on its features (i.e. tweet words). Of the various classification techniques, Bayes and SVM **[14]** are examined in this study.

Naive Bayes is a supervised machine learning algorithm that can be trained quickly and also make fast prediction. It assumes that the occurrence of a event is not dependent on the occurrence of another event. The Bayes Theorem states that the probability of event B given A is equal to the probability of event A given B multiplied by probability of A divided by probability of B. The expression for probability of event A when event B is true,

$$P(A/B) = \frac{P(B/A)*P(A)}{P(B)} \quad (1)$$

where,

$P(A/B)$ = probability of occurrence of event A

$P(A)$

$P(B/A)$ = probability of occurrence of event B given A is true

Multinomial Naive Bayes is a specialized version of Naive Bayes which is designed mainly for text documents [25]. In this study, the Naive Bayes classifier by Textblob [21] was used to choose a label ("Related" or "notRelated").

Support Vector Machine is a supervised learning method that analyzes the data and recognizing patterns used for classification **[14]**. Support Vector Machines rely on a hyper-plane isolating plane to define a classifier. SVM is defined by a separating hyperplane that maximizes the margin between the two classes and minimize the empirical classification error. Optimum hyperplane search function shown in (2) subject to (3). Scoring in SVM to determine the test document class use (4). In this study, scikit-learn **[20]** implementation of linear SVM is used to classify whether the tweets are related to Nepal tourism or not.

$\frac{1}{2}w^T$w+C$\Sigma_i\xi_i$ (2)

$(x_i,y_i),y_i(w^T x_i+b) \geq 1-\xi_i$ (3)

where,

w=weight vector, C=loss function,
ξ_i =slack variable/misclassification vector i,
x_i=train vector i, y_i=class train vector i,
$b = biasvalue$

$f(x)$=sign(w^Tx+b) (4)

where,

$f(x)$=score function,x=vector test document

4. Results & Discussion

In this section, we present the results. First, we discuss about the tweet dataset and then about the performance of the classification algorithms on the tweet classification problem.

4.1. Tweet Dataset

For the analysis, keywords related to Nepal from popular Nepal tourism related sources like Nepal Tourism Board, www.toppkr.gov.np, www.welcomenepal.com, www.touropia.com, etc. were obtained. Next, tweets containing those keywords were collected. Further, tweets not related to Nepalese tourism were also gathered. Next, two tourism domain people independently examined the tweets and finally a total of 200 tweets were collected; 100 tweets were related to Nepalese tourism and the rest related were not related to Nepalese tourism. From that dataset, 50% of the data is used as training set and the remaining 50% of the data were used as for testing purpose. A snapshot of the portion of the training and testing dataset is shown in Table 1 and Table 2.

4.2. Classification Report of Naive Bayes and SVM Classifier

As per the objectives of the research, we classified the tweets using Naive Bayes and Support Vector Machine. A summary of the tweet classification report is shown in Table 3 and Table 4.

Figure 3 illustrates a plot which shows the comparison of precision, recall, F1 score and accuracy of SVN and Naive Bayes algorithm. As per the result, the overall accuracy and F1 score of Naive Bayes algorithm is 81% and 0.91 respectively. Support Vector Machine algorithm showed better performance with the overall accuracy of 91% and F1 Score of 0.91. Similarly, SVM outperformed Bayes in terms of overall precision and recall as well. Hence, for these kinds of problems, it indicates that SVM is more desirable than Naive Bayes algorithm.

Table 1: Some data from Training Set

#Pokhara is Nepals best #adventure and #leisure city. Amazing Pictures.	Related
Beautiful #Pokhara	Related
We brought clean water to more than 100 families suffering from last years earthquakes.	notRelated
The ship is beautiful addition to the #Portsmouth skyline	NotRelated
A spectacular view of majestic Mt #Machhapuchhre in #Pokhara	Related

Table 2: Some data from Testing Set

Swet vairab at Basantapur durbar square during #indrajatra festival.	Related
#Neapl so rich in culture that everyone gets mesmerized by it. #Mask dance in #Bhakrapur	Related
I got lost in Amsterdam again today but this time with @RealKiraleelee	notRelated
Use Machine Learning at sale to built better products.	NotRelated
Throwback to when the Himalayan squad made it to Annapurna Base Camp	Related

Table 3: Classification Report for SVM

Class	Precision	Recall	F1 Score	Accuracy
Related	1	0.82	0.9	91%
notRelated	0.85	1	0.92	
Overall	0.92	0.91	0.91	

Table 4: Classification Report for Bayes

Class	Precision	Recall	F1 Score	Accuracy
Related	1	0.62	0.77	81%
notRelated	0.72	1	0.84	
Overall	0.86	0.81	0.8	

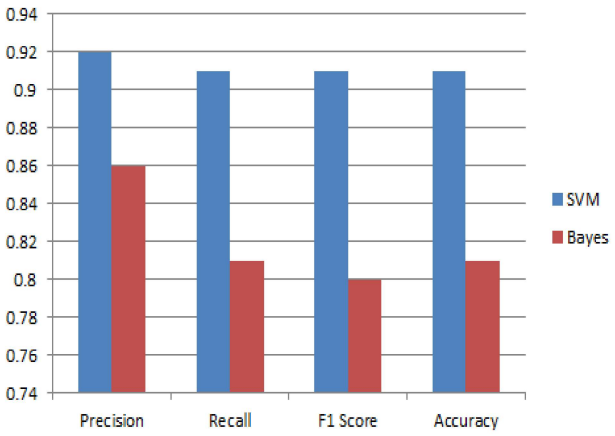

Figure 3: Comparison graph of SVM and Naive Bayes Classifier

4.3. Discussion of the Result obtained from Naive Bayes and SVM Classifier

Naïve Bayes classifier is based on the hypothesis that each feature is mutually independent. It is simple, easy to use and generally shows good efficiency. However, compared with other algorithms it may not perform on par in text classification tasks. It may be due to its inherent assumption of independent attributes, text redundant features and rough parameter estimation [29]. SVMs, in theory, acknowledge the inherent characteristics of text like high dimensional feature spaces and sparse instance vectors [30]. Because of their ability to automatically tune acceptable parameter settings and generalize effectively in high-dimensional feature spaces, SVMs reduce the requirement for feature selection, making text categorization much easier to implement. Most importantly, the most significant distinction in terms of "features" is that Naive Bayes assumes them as independent, whereas SVM considers their interactions to some extent. As a result, SVM provides better text categorization results in our experiment.

This investigation implies utilizing automatic text categorization can be implemented to detect tourism related content in online social networks like Twitter. This can act as an important source for revealing various tourism aspects in the real-time. The outcomes of this study can provide valuable insights regarding the Nepalese Tourism Brand presence in social media and can act as an starting point for further in depth investigations. It can provide a valuable reference for various stakeholders such as tourism planners, urban planners, and so on.

5. Conclusions and Future Work

In this paper we examined the problem of tweet classification in order to automate the process of identifying Nepalese tourism related contents. We have collected tweets and developed a dataset of tourism related and non-related tweets. We used two popular text classification algorithms and accomplished experiments on manually collected tweet data set. Naive Bayes and SVM algorithm were examined and their performance was accessed. The significant part of this study is automating the identification of Nepalese tourism tweets by applying supervised classification methods and then further evaluating the performance of such classification algorithms. The performance measures obtained from applying the selected algorithms reveal that SVM algorithm outperformed Naive Bayes algorithm. The overall

values for precision, recall, F1 score and accuracy of SVM is better than Naive Bayes. F1 Score of 0.91 and 0.80 is obtained respectively for SVM and Bayes algorithm. Similarly, average accuracy of the results of SVM and Naive Bayes are 91% and 81% respectively. The outcomes of this study can provide valuable insights regarding the Nepalese Tourism Brand presence in social media and can act as an starting point for further indepth investigations. It can provide a valuable reference for various stakeholders such as tourism planners, urban planners, and so on.

In the future we are looking forward to enlarge the Nepal Tourism related Twitter dataset and apply classification algorithms at a larger scale. Further improvement can be done by using native Nepalese language for classification (not just English language tweets). Also, we plan to perform foreigner tweet analysis as well as Nepalese tweet analysis by focusing on various aspects of tourism.

References

[1] Kim W., Jeong O. and Lee S. On social Web sites", (2009) [Online]. Available: http://dx.doi.org/10.1016/j.is.2009.08.003.

[2] Social Networking Statistics, (2015) [online]. Available: http://www.statisticbrain.com/social-networking-statistics/

[3] "Twitter. Inc , (2020) [Online]. Available: https://twitter.com/.

[4] "Alexa Top 500 Global Sites", *Alexa.com*, (2017) [Online]. Available: http://www.alexa.com/topsites. [Accessed: 22- May- 2017].

[5] STATS | Twitter Company Statistics - Statistic Brain, *Statistic Brain*, (2017). [Online]. Available: http://www.statisticbrain.com/twitter-statistics. [Accessed: 21- May- 2017].

[6] Lee W. J., Oh K. J., Lim C. G. and Choi H. J. User profile extraction from Twitter for personalized news recommendation, *Advanced Communication Technology (ICACT), 2014 16th International Conference on, Pyeongchang*, (2014) 779-783. doi: 10.1109/ICACT.2014.6779068

[7] Li J., Ritter A., Cardie C., and Hovy E. Major life event extraction from twitter based on congratulations/condolences speech acts. *In Proceedings of Empirical Methods in Natural Language Processing*, (2014).

[8] Wang X., Zhu F., Jiang J., Li S. Real time event detection in twitter, *Proceedings of the 14th international conference on Web-Age Information Management*, Beidaihe, China, (2013).

[9] Goeldner, C., Ritchie, B., and McIntosh. R. (2000). "Tourism: principles, practices and philosophies (8th edition)". New York: John Wiley and Son

[10] Cathy Schetzina, The PhoCusWright Consumer Technology Survey Second Edition, *Trends and Issues in Global Tourism 2009*, (2009) 113-133, DOI10.1007/978-3-540-92199-8_8

[11] Location, Location, Location, https://blog.twitter.com/2009/location-location-location.

[12] Adding your location to a Tweet, (2015) [Online].Available: https://support.twitter.com/articles/122236

[13] VISIT NEPAL 1998, (2015). [Online]. Available: www.travel-nepal.com/vny98/

[14] C.D. Manning, P. Raghavan, H. Schutze. Introduction to Information Retrieval. Cambridge UP, (2008)

[15] Jalal S. Alowibd, Ghani S., Mokbel M., VacationFinder: A Tool for Collecting,

Analyzing, and Visualizing Geotagged Twitter Data to Find Top Vacation Spots, *7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN)* (2014) Dallas, Texas.

[16] Oku K., Hattori F. Mapping Geotagged Tweets to Tourist Spots Considering Activity Region of Spot, *Tourism Informatics*, (2015). DOI 10.1007/978-3-662-47227-9_2

[17] Shimada K., Inoue S., Maeda H. and Endo T. Analyzing Tourism Information on Twitter for a Local City, *Software and Network Engineering (SSNE), 2011 First ACIS International Symposium on, Seoul*, (2011) 61-66. doi:10.1109/SSNE.2011.27

[18] Shimada K., Inoue S., Maeda H. and Endo T. On-site Likelihood Identification of Tweets for Tourism Information Analysis, *Advanced Applied Informatics (IIAIAAI), 2012 IIAI International Conference on, Fukuoka*, (2012) 117-122. doi:10.1109/IIAI-AAI.2012.32

[19] Devkota, B. Online Social Networks and its potentialities for Nepalese Tourism Promotion, *OODBODHAN*, 4 (1) (2016).

[20] Scikit-learn Project, (2020). Scikit-learn. https://scikit-learn.org/

[21] TextBlob Project, (2020). TextBlob. https://textblob.readthedocs.io/en/dev/

[22] Devkota, B., and Hiroyuki M. An Exploratory Study on the Generation and Distribution of Geotagged Tweets in Nepal. *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS). IEEE*, (2018).

[23] Devkota, B., Hiroyuki M., and Pahari N. Utilizing User Generated Contents to describe Tourism Areas of Interest, *First International Conference on Smart Technology & Urban Development (STUD 2019) December 13-14*, Chiang Mai, Thailand (2019).

[24] Devkota, B. et al. Using Volunteered Geographic Information and Nighttime Light Remote Sensing Data to Identify Tourism Areas of Interest. *Sustainability* 11.17 (2019): 4718.

[25] McCallum A. and Nigam K. A Comparison of Event Models for Naive Bayes Text Classification.

[26] Zubrinic K., Milicevic M. and Zakarija I. Comparison of Naive Bayes and SVM Classifiers in Categorization of Concept Maps. *International journal of computers,* 7 (2013) 109-116.

[27] Popa I.S., Zeitouni K., Gardarin G. Text Categorization for Multi-label Documents and many Categories, *Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07)*, (2007).

[28] Sebastiani F., Machine Learning in automated text categorization, *ACM Comput. Surveys*, 34 (2002).

[29] Joachims T. Text categorization with support vector machines: Learning with many relevant features. *InEuropean conference on machine learning 1998 Apr 21,* (1998) 137-142. Springer, Berlin, Heidelberg.

[30] Chen H., Fu D. An improved Naive Bayes classifier for large scale text. *In2018 2nd International Conference on Artificial Intelligence: Technologies and Applications (ICAITA 2018) (2018) Atlantis Press,* 33-36.