Celebrity Profiling from Twitter Data

Kumar Gourav Das[1], Braja Gopal Patra[2] and Sudip Kumar Naskar[3]
[1]*Department of Computer Science & Engineering, Future Institute of Engineering & Management, Kolkata, India*
[2]*Department of Biostatistics and Data Science, School of Public Health,*
University of Texas Health Science Center, Houston, TX, USA
[3]*Department of Computer Science & Engineering, Jadavpur University, Kolkata, India*
Email: kumargouravdas18@gmail.com, brajagopal.cse@gmail.com, sudip.naskar@gmail.com

***Abstract*—Twitter is one of the most popular social media platforms which enables users from different walks of life to have two-way communications. User categorization in Twitter platform categorizes a group of users based on their profiles, posts, and tweeting behaviors. The purpose of user categorization is to deliver relevant information to a specific class of users based on their interests. In this work, we perform user-level categorization of celebrities based on their professions. The accuracy of the proposed model is evaluated for three different datasets(Indian,non-Indian and Combined) using various machine and deep learning frameworks. In this regard four different light weight features have been extracted(stylistic, hashtag, embeddings, and topic-based features) from the dataset and Convolutional Neural Network is the best performing model for all three datasets. The maximum accuracies obtained are 76.66%, 91.72% and 83.01% for Indian, non-Indian, and combined datasets, respectively.**

***Keywords*-Social media; celebrity profiling; embedding features; convolutional neural network;**

I. Introduction

Social media platforms are web-based communication applications that enable users to interact with each other by both sharing and consuming information [1]. Existing social media platforms such as Facebook, Twitter, LinkedIn, and WhatsApp, have attracted widespread research in terms of data analysis over widely generated user contents, recommendations [2], referral systems, security [3], communication, as well as computing perspectives in both mobile and web-based heterogeneous platforms.

Social media like Twitter has given users a closer look at their favorite celebrities, allowing them directly communicate with the celebrities , which influences their fans. It is essential to classify the celebrities into categories which helps to deliver the relevant contents to the interested users/fans. Celebrity profiling will help in social recommendation where users are recommended informative contents/topics from similar celebrities rather than providing with unnecessary information. For example, a Twitter user interested in witty and humorous contents and following James Corden in Twitter (but unaware about Michael McIntyre) is also very likely to follow Michael McIntyre if the user is recommended about Michael McIntyre.

Many experiments have already been performed in the area of social media text analytics. However, there are some crucial aspect to these research studies that remains unexplored. Social media users have their own different choices of interests. Thus recommending a celebrity to a user from ones list of interest is essential , otherwise it may lead to an unnecessary flow of data across users, which increases the overall complexity of the system and diminishes the efficacy of the platform. Further, celebrity profiling may help in suggesting Twitter accounts to follow in addition to other information such as location, activity on Twitter[1]. To this end, classification techniques are highly useful for categorizing Twitter celebrities as per the objective of this proposed work.

User profiling is an established problem in natural language processing (NLP) domain. The authors propose celebrity profiling as a niche domain of user profiling. Celebrity profiles often generate mass interest, and their posts can influence followers and society in general. Celebrity profiling is a research area which can be associated with credible knowledge sources and used for training an efficient computational algorithm for categorizing celebrities into different groups [4]. The work presented in this paper aims to classify celebrity profiles into different predefined categories using various textual features. The contributions in this paper are threefold: First, collecting a celebrity profile dataset from Indian and non-Indian celebrities. Second, exploring various features for classifying celebrities into six predefined categories. Third, developing different celebrity category classification systems using machine and deep learning algorithms.

This paper is organized as follows: Section II surveys the state-of-the-art in author and celebrity profiling tasks. Section III describe the prepared celebrity profiling dataset and features used in this experiment. Results of the systems and detailed discussion are mentioned in Section IV. Finally, the conclusion and future direction are drawn in Section V.

II. Related Work

In recent times, social media platform like Twitter has millions of users; thus classification of Twitter user is essential, because it has several important applications in advertising, personalizing, and recommendation system. PAN is a series of shared tasks on plagiarism, author profiling, and other related tasks. Recently, PAN organized celebrity profiling [4], [5] for the first time and the task focuses on identifying fame, occupation, age,

[1] https://help.twitter.com/en/using-twitter/account-suggestions

978-1-7281-5014-7/19/$31.00 ©2019 IEEE

and gender of celebrities from their tweets[2]. Here, we mention top performed systems of some participants in the PAN celebrity profiling task. Support vector machines and logistic regression based systems using TF-IDF of word and character n-grams by Radivchev et al. [6] obtained top performance in the task. Sandoval et al. [7] used logistic regression with n-gram and linguistic features to predict age, gender, and fame. They used a multinomial naive Bayes model with n-gram and linguistic features to predict occupation. Martinc et al. [8] used TF-IDF vectors and trigrams of the first 100 tweets per timeline to predict all four demographics using logistic regression. Other than these, there are no other publications related to celebrity profiling/categorization. A significant amount of research has been performed in the field of author profiling in tweets. Different author profiling tasks have been discussed below.

The author profiling task includes identifying several aspects of users such as gender, age group, native language, personality type, etc. from texts [9], [10]. Further, multimodal author profiling has been performed based on tweets and images associated with tweets using different machine and deep learning algorithms [11], [12], [13]. Raghuram et al. [14] categorized 593 Indian twitter users into six different classes (Politics, Health care, etc.) using 500 most recent tweets of them.

In another contribution, De Choudhury et al. [15] categorized users as organizations, journalists/media bloggers, and ordinary individuals based on tweets. The task was performed using k-NN on 4,932 twitter users containing 200 most recent tweets of each user. In another notable work, Pennacchiotti and Popescu [16] built a machine learning model that identifies the user's ethnicity, political affiliation, and whether a user is attracted towards a particular brand of business or not. The model performed better in the latter two tasks.

Rao et al. [17] investigated gender, age, regional origin, and political orientation using socio-linguistic and n-gram features. Ikeda et al. [18] proposed demographic estimation algorithm that tries to identify the user's age, gender, location, hobby, occupation, and marital status from tweets and community relationships. Hoang et al. [19] predicted locations of the twitter user from their tweets.

III. Data and Feature Extraction

A. Data

The dataset consists of tweets from both *Indian* and *Non-Indian* celebrities and the dataset was collected using tweepy API[3]. Due to the restriction of tweepy API, most recent 3,200 tweets for every individual user has been collected during the time span of July 27, 2017, to August 2, 2018.

For both *Indian* and *non-Indian* celebrities, the dataset consists of six different classes according to professions of twitter users such as Business (BUS), Education (EDU), Entertainment (ENT), Politics (POL), Sports (SPRT) and Technology (TECH). We did not merge Technology and Business classes because we have plans to classify celebrities with multiple classes (multi-label classification problem) in the future. Each tweet must have the following three fields, Tweet ID, Date & time, and tweet. The overall statistics of the dataset based on their gender has been shown in Table I. The distribution of male and female users is not uniformed in the dataset. In the future, size of the dataset may be increased, and uniformity among male and female users can be maintained.

Table I: Gender wise details of the collected dataset

Classes	Indian		non-Indian	
	Male	Female	Male	Female
BUS	86	7	72	27
EDU	71	25	66	48
ENT	68	33	54	84
POL	90	9	70	28
SPRT	90	11	104	17
TECH	68	21	124	26
Total	473	106	490	230

Celebrity level manual annotation has been performed on the collected dataset after reading tweets of each celebrity. We faced difficulties while annotating celebrities from Business, Education, and Technology categories. Celebrities from these three classes often belong to multiple classes and therefore, these three classes are not disjoint in reality. In future, we have a plan to annotate categories at tweet level and more detailed classification for each category can be performed. In this work, celebrities were categorized into separate categories. However, more fine-grained classification can be performed in future (e.g., classifying a sports person into a cricketer or a footballer).

Many non-English tweets/words are written in English alphabet (i.e., code-mixed tweets). Indian celebrities tend to write in code-mixed languages. Hence, the non-English tweets were removed using an English lexical database, where each tweet was tokenized and lemmatized using *nltk* [20]. Then lemmatized word forms were checked for presence in the dictionary. Tweets with more than 50% non-English words were removed from further processing [21].

The details of the dataset used in the experiments are provided in Table III, which include the number of users, total number of tweets, maximum, average and minimum number of tweets for each class,and average tweet rate for each class. The average tweet rate is the average time difference in minutes between two successive tweets for each celebrity in each class. Average tweet rate is calculated on collected tweets. Some instance of username from every class for both Indian and non-Indian users has been shown in Table II.

Preprocessing: The data cleaning process was performed in the following steps: all the non-ASCII characters (such as ©, ®, ^{TM}, etc.), URLs, hyperlinks, '@' and special characters were removed.

[2] https://pan.webis.de/clef19/pan19-web/celebrity-profiling.html

[3] https://www.tweepy.org/

Table II: Some examples of accounts of Indian and non-Indian Twitter celebrities

Class	Indian	non-Indian
BUS	@AnupamMittal, @anandmahindra, @kishore_biyani, @MPNaveenJindal.	@marshawright, @PamMktgNut, @timoreilly, @ckburgess.
EDU	@PrakashJavdekar, @Meetasengupta, @PKumar59, @rameshmashelkar.	@CarolCampbell4, @AngelaMaiers, @CreativeSTAR, @Doug_Lemov.
ENT	@SrBachchan, @karanjohar, @ashabhosle, @AnilKapoor.	@Scarlett_Jo, @shakira, @Pitbull, @channingtatum.
POL	@arunjaitley, @AmitShah, @MamataOfficial, @jadeja_mp.	@PutinRF_Eng, @stephenharper, @PeterDutton_MP, @realDonaldTrump.
SPRT	@suvodipmoitra, @sachin_rt, @Pvsindhu1, @imVkohli	@BrettLee_58, @AllanDonald33, @KumarSanga2, @MesutOzil1088.
TECH	@AbhijitBhaduri, @sundarpichai, @satyanadella,@RPrasad12	@jeffweiner, @jordanrcrook, @LanceUlanoff, @jasonhowell.

Table III: Statistics of the celebrity tweet datasets (CAT: categories, NC: number of celebrities, NT: number of tweets, MAX: maximum number of tweets, AVG: average number of tweets, MIN: minimum number of tweets, ATR: average tweet rate - average time difference in minutes between two successive tweets)

	CAT	NC	NT	MAX	AVG	MIN	ATR
Indian	**BUS**	93	224,569	3200	2415	119	2.56
	EDU	96	227,730	3200	2372	318	4.04
	ENTR	101	289,332	3200	2865	118	3.89
	POLY	99	234,743	3200	2371	187	4.23
	SPRT	101	179,482	3200	1777	207	3.21
	TECH	89	208352	3200	2341	127	3.60
	TOTAL	579	1,364,208	NA	NA	NA	NA
non-Indian	**BUS**	99	299,683	3200	3027	185	3.80
	EDU	114	347,867	3200	3051	458	3.07
	ENTR	138	306,993	3200	2225	238	5.80
	POLY	98	289,979	3200	2959	550	4.28
	SPRT	121	309,591	3200	2559	258	6.41
	TECH	150	387,592	3200	2584	248	4.32
	TOTAL	720	1,941,705	NA	NA	NA	NA

B. Feature Extraction

Feature extraction plays a pivotal role in any machine learning framework. Different content-based features extracted from profiles and tweets of celebrities were used in our experiments. A detailed description of the features used is mentioned below.

Stylistic: The number of stopwords, punctuation, smileys, tweets or re-tweets, hashtags, and slang words were considered in the present study. The stopword list is collected from *nltk corpus*[4]. The slang word list was manually prepared. Emoji package[5] was used for identifying the emoji in the tweet and then counted the frequency of such smileys. Details of all Stylistics features and their description are shown in Table IV. The stylistic features were collected before preprocessing steps.

Word embedding: Word embedding gained immense popularity in the last few years and it has been successfully used in several text mining and NLP tasks. It is in fact an inseparable component of every deep learning based state-of-the-art NLP systems. For the present study, word vector representations were obtained using the word2vec model, *GloVe* [22]. *Glove* is pretrained on 2 billion tweets, and

Table IV: Stylistics features of tweets used in our experiments

Features	Description
Stop-words	Total number of stopwords in tweets of a user.
Repeat Punctuation	How many times a punctuation repeats.
Punctuation	Total number of punctuation present in tweets of a user.
Happy Emoticons	Total number of happy emoticons of a user's tweets.
Sad Emoticons	Total number of sad emoticons of a user's tweets.
Emoji	Total number of Emoji of a user's tweets.
Hash Count	Total number of hashtag present in tweets of a user.
Slang Word	Total number of slang words in tweets of a user.
Direct Tweet	Total number of direct tweets (@) of each user.

[4] https://www.nltk.org/nltk_data/

[5] https://pypi.org/project/emoji/

it provides dimensional flexibility. *GloVe* delivers a single feature vector for each word and the vector representation of a tweet is constructed from the vector representation of its constituent words as in Equation 1. Finally, celebrity vector is constructed as in Equation 2 by adding together the tweet vectors of his/her tweets.

$$\overrightarrow{t_i} = \frac{1}{n_i}\sum_{j=1}^{n_i}\overrightarrow{w_{ij}}, \forall w_{ij} \in GloVe \qquad (1)$$

$$\overrightarrow{C_k} = \frac{1}{N_k}\sum_{i=1}^{N_k}\overrightarrow{ti} \qquad (2)$$

Here, $\overrightarrow{t_i}$ denotes the tweet vector for the i^{th} tweet; $\overrightarrow{w_{ij}}$ represents the j^{th} word vector of $\overrightarrow{t_i}$ and n_i is the number of words present in $\overrightarrow{t_i}$; $\overrightarrow{C_k}$ refers to the celebrity vector of the k^{th} celebrity and N_k is the number of tweets posted/shared by $\overrightarrow{C_k}$. Word vectors of dimension 200 were used in our experiments.

Topic words: It is useful to collect the topic words which can describe the whole document in a few words. We used Latent Dirichlet Allocation (LDA) [23] model implemented in gensim [24] to collect three important topics containing ten words each, for a single celebrity. The resulted output was converted into 100-dimensional vector.

Hashtags: Hashtags are informative content on microblogs like Twitter. The extensive use of hashtags in the dataset motivated us to use it as a feature. We used Latent Semantic Analysis (LSA) [25] to compute a 25-dimensional feature vector from the hashtags for each celebrity.

IV. Results and Discussion

A series of experiments have been conducted using different machine and deep learning frameworks. The experiments are performed after categorizing celebrities into three segments namely i)*Indian*, ii)*non-Indian* and iii) *combined*. Effectiveness of different features has been assessed using four different classification models - Support vector machine (SVM), Naïve Bayes (NB), Multilayer Perceptron (MLP) and Convolution Neural Network (CNN).

We used a 10-fold cross-validation framework for our experiments. Table V lists the classification accuracies for the three datasets on an individual as well as combined features. The CNN model provides the overall best classification accuracies across all the three datasets (76.66, 91.72 and 83.01 for the *Indian*, *non-Indian* and *combined* datasets, respectively) on the combined features. In Table V, feature-specific best scores are shown in italics, while the model-specific best scores are underlined. A comparison of the model-specific best scores (the underlined ones) reveals that word embeddings are the most effective features for SVM, MLP, and CNN, while NB works best with topic features.

It may be noted that the accuracies on the *Indian* dataset are less than others. This is primarily because the tweets of the *Indian* celebrities are mostly code-mixed as observed from the dataset. Thus many null vectors ($\overrightarrow{0}$) were generated by *GloVe* for the unseen words, which resulted in degradation of accuracy for the Indian and combined datasets. It was also observed from the results that the system was not able to properly distinguish between the celebrities in Education, Business and Technology domains; this finding is also in sync with the annotation difficulties mentioned in Section III-A. Category-specific accuracies of *Indian*, *non-Indian* and *combined* datasets using different machine and deep learning algorithms are shown in Table VI. The category-specific best scores for each dataset are marked as bold.

In Raghuram et al. [14], the authors categorized Indian Twitter users into Politics, Entertainment, Entrepreneurship, Journalism, Science & Technology, and Health care using different machine learning frameworks. However, the present work is more focused on celebrity profiling. The number of users and their tweets are also considerably high in the work presented here. Recently, PAN organized celebrity profiling task and the goal of this task is to identify gender, age, fame, and occupation. We mention some of the top-performing systems next. The system of Radivchev et al. [6] predicts age, gender, fame, and occupation with 93%, 51%, 77%, and 75% accuracies, respectively. The system submitted by Moreno-Sandoval et al. [7] predicts gender, age, fame, and occupation with 86%, 37%, 54%, and 72% accuracies, respectively. The system submitted by Martinc et al. [8] obtains 91%, 44%, 75%, and 73% accuracies for predicting gender, age, fame, and occupation, respectively. However, it is difficult to compare the proposed methods in this paper with the systems submitted to PAN celebrity profiling tasks because the dataset is totally different from the dataset that we collected. In the near future, we plan to implement our methods on PAN celebrity profiling dataset.

V. Conclusion and Future Work

This paper proposes a complete classification framework for automatic Twitter celebrity classification into six categories, namely Business, Education, Entertainment, Politics, Sports, and Technology. Three large annotated datasest for celebrity profiling has been created. Multiple features extracted from these three datasets are further incorporated into several machine and deep learning classifiers for comparative performance evaluation. CNN is the best performing classifier for all three datasets on combined feature set.

Initially, we started our work with these six categories. This can be further extended to more categories in future. We are planning to include more categories like Health Care, Journalism, etc. Another viable option is to increase the size of the datasets by considering more celebrities for each class. However, this process may consume more time as the data collection mechanism is a resource-intensive as well as time-consuming task. This datasets may be annotated and used later for gender, age group, ethnicity classification, etc.

Among the six classes, classifying celebrities from Business, Education, and Technology categories is chal-

Table V: Classification results produced by four different models on three different datasets on different feature sets

		Features				
Celebrities	**Classifiers**	**Stylistic**	**Hashtags**	**Embeddings**	**Topics**	**Combined**
Indian	**NB**	32.20	29.15	59.83	63.89	62.71
	SVM	41.13	37.83	73.22	*72.06*	71.42
	MLP	41.69	*57.79*	74.91	70.16	76.61
	CNN	*46.66*	42.50	*75.00*	68.33	***76.66***
non-Indian	**NB**	26.53	13.88	71.94	78.89	78.19
	SVM	*49.30*	36.01	82.82	79.72	81.83
	MLP	48.05	*56.94*	87.50	80.55	89.72
	CNN	20.03	44.13	*88.96*	*88.34*	***91.72***
Combined	**NB**	28.09	28.74	55.50	60.35	61.97
	SVM	*35.27*	30.99	74.67	68.77	75.56
	MLP	35.10	*48.57*	*80.06*	*74.13*	76.35
	CNN	34.81	30.94	78.49	71.32	**83.01**

Table VI: Category wise classification accuracies of four different classifiers on three datasets using all combined features. IN: Indian, NIN: non-Indian, COM: Combined

Classes	**NB**			**SVM**			**MLP**			**CNN**		
	IN	**NIN**	**COM**	**IN**	**NIN**	**COM**	**IN**	**NIN**	**COM**	**IN**	**NIN**	**COM**
BUS	17.39	58.58	19.79	57.60	74.74	68.06	**65.21**	75.75	**73.82**	45.00	**85.00**	72.50
EDU	44.32	80.70	65.40	72.16	75.00	78.28	64.94	94.73	**82.46**	**75.00**	**95.65**	79.06
ENT	80.39	86.23	88.33	92.15	90.57	91.66	91.16	93.47	90.83	**95.00**	**96.42**	**97.91**
POL	76.76	79.59	51.26	91.91	89.79	80.20	93.93	90.81	**89.34**	**95.00**	**96.90**	87.50
SPRT	89.88	84.29	55.65	84.00	96.69	80.09	85.00	96.69	85.52	**95.00**	**97.30**	**90.90**
TECH	**82.02**	74.66	80.33	69.66	**86.00**	75.73	65.16	85.33	**77.40**	61.00	73.33	72.91

lenging because it has been observed that the interests of these celebrities vary among different categories. For example, a celebrity from the business category may have an interest in politics. Thus, multi-class classification of celebrities can be an immediate future task. In this paper, celebrity level classification was performed; however, individual celebrity's tweet level analysis can provide more fine-grained analysis. Given a large amount of code-mixed celebrity tweets, the future extension of the proposed work may include code-mixed celebrity profiling.

References

[1] B. Batrinca and P. C. Treleaven, "Social media analytics: a survey of techniques, tools and platforms," *Ai & Society*, vol. 30, no. 1, pp. 89–116, 2015.

[2] M. Pennacchiotti and S. Gurumurthy, "Investigating topic models for social media user recommendation," in *Proceedings of the 20th international conference companion on World wide web*. ACM, 2011, pp. 101–102.

[3] I. Inuwa-Dutse, M. Liptrott, and I. Korkontzelos, "Detection of spam-posting accounts on twitter," *Neurocomputing*, vol. 315, pp. 496–511, 2018.

[4] M. Wiegmann, B. Stein, and M. Potthast, "Celebrity profiling," in *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 2019, pp. 2611–2618.

[5] ——, "Overview of the celebrity profiling task at pan 2019," in *Working Notes Papers of the CLEF*, 2019.

[6] V. Radivchev, A. Nikolov, and A. Lambova, "Celebrity profiling using tf-idf, logistic regression, and svm notebook for pan at clef 2019," in *Working Notes Papers of the CLEF*, 2019.

[7] L. G. Moreno-Sandoval, E. Puertas, F. M. Plaza-del Arco, A. Pomares-Quimbaya, J. A. Alvarado-Valencia, and L. A. Ureña López, "Celebrity profiling on twitter using sociolinguistic features notebook for pan at clef 2019," in *Working Notes Papers of the CLEF*, 2019.

[8] M. Martinc, B. Škrlj, and S. Pollak, "Who is hot and who is not? profiling celebs on twitter notebook for pan at clef 2019," in *Working Notes Papers of the CLEF*, 2019.

[9] B. G. Patra, S. Banerjee, D. Das, T. Saikh, and S. Bandyopadhyay, "Automatic author profiling based on linguistic and stylistic features notebook for pan at clef 2013," in *Working Notes Papers of the CLEF*, 2013.

[10] F. Rangel, F. Celli, P. Rosso, M. Potthast, B. Stein, and W. Daelemans, "Overview of the 3rd author profiling task at pan 2015," in *CLEF 2015 Evaluation Labs and Workshop Working Notes Papers*, 2015, pp. 1–8.

[11] G. Farnadi, J. Tang, M. De Cock, and M.-F. Moens, "User profiling through deep multimodal fusion," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 2018, pp. 171–179.

[12] B. G. Patra, K. G. Das, and D. Das, "Multimodal author profiling for twitter notebook for pan at clef 2018," in *Working Notes Papers of the CLEF*, 2018.

[13] F. Rangel, P. Rosso, M. Montes-y Gómez, M. Potthast, and B. Stein, "Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter," in *Working Notes Papers of the CLEF*, 2018.

[14] M. Raghuram, K. Akshay, and K. Chandrasekaran, "Efficient user profiling in twitter social network using traditional classifiers," in *Intelligent systems technologies and applications*. Springer, 2016, pp. 399–411.

[15] M. De Choudhury, N. Diakopoulos, and M. Naaman, "Unfolding the event landscape on twitter: classification and exploration of user categories," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 2012, pp. 241–244.

[16] M. Pennacchiotti and A.-M. Popescu, "Democrats, republicans and starbucks afficionados: user classification in twitter," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 430–438.

[17] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*. ACM, 2010, pp. 37–44.

[18] K. Ikeda, G. Hattori, C. Ono, H. Asoh, and T. Higashino, "Twitter user profiling based on text and community mining for market analysis," *Knowledge-Based Systems*, vol. 51, pp. 35–47, 2013.

[19] T. B. N. Hoang, V. Moriceau, and J. Mothe, "Predicting locations in tweets," in *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CLICLing 2017)*, 2017.

[20] E. Loper and S. Bird, "Nltk: the natural language toolkit," *arXiv preprint cs/0205028*, 2002.

[21] S. Mandal, M. Rath, Y. Wang, and B. G. Patra, "Predicting zika prevention techniques discussed on twitter: An exploratory study," in *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. ACM, 2018, pp. 269–272.

[22] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[24] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010, pp. 45–50.

[25] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.