Spatial-based Topic Modelling using Wikidata Knowledge Base

Kwan Hui Lim*, Shanika Karunasekera*, Aaron Harwood*, and Lucia Falzon†*
*School of Computing and Information Systems, The University of Melbourne, Parkville, Victoria, Australia
†Defence Science and Technology Group, Edinburgh, South Australia, Australia
Email: {kwan.lim, karus, aharwood}@unimelb.edu.au, Lucia.Falzon@dst.defence.gov.au

***Abstract*—Topic modelling is a well-studied field that aims to identify topics from traditional documents such as news articles and reports. More recently, Latent Dirichlet Allocation (LDA) and its variants, have been applied on social media platforms to model and study topics relating to sports, politics and companies. While these applications were able to successfully identify the general topics, we posit that standard LDA can be augmented with spatial and temporal considerations based on the geo-coordinates and timestamps of social media posts. Towards this effort, we propose a spatial and temporal variant of LDA to better detect more specific topics, such as a particular art exhibit held at a museum or a security incident happening on a particular day. We validate our approach on a Twitter dataset and find that the detected topics are well-aligned to real-life events happening on the specific days and locations.**

***Index Terms*—Topic Modelling, Latent Dirichlet Allocation, Twitter, Microblogs**

I. Introduction and Related Work

Twitter is a popular micro-blogging platform that allows users to share information and news with their friends, in the form of short messages of up to 140 characters. Apart from its usage as a social networking platform, Twitter has also been used for other purposes, such as predicting disease outbreaks [1], managing crises and emergencies [2], studying political conversations [3], among others. One key challenge in these applications is the detection of a relevant topic being discussed in a tweet, and a popular approach is to apply the well-known Latent Dirichlet Allocation (LDA) [4] to determine the topics discussed in these tweets. LDA is a topic model that probabilistically assigns each document (tweet) to a set of multiple topics, where each topic is represented by a set of words.

LDA is a popular algorithm that has been frequently used to model topics in Twitter, and we now review the main works in this area. Researchers have performed empirical studies using variants of LDA on Twitter, where the variants are based on different tweet aggregation schemes, e.g., by author, terms, hashtags, and others [5], [6]. Similarly, researchers [7], [8] have also utilized community detection algorithms on various tweet-based networks to identify the main discussion topics. Others like [9] utilized a variant of LDA to study topical differences between content in Twitter and New York Times. LDA has also been used for trending topic detection, such as [10] who used variants of LDA with different pre-processing steps to identify trending topics relating to the Football Association Challenge Cup and US Elections. On the same note, [11] also studied trending topics relating to commercial companies, focusing on identifying the onset of topic discussion to the time of wide-spread trending.

While LDA has been successfully used to study Twitter data, many of these applications aim to study general topics without considering the time and place where the topics were discussed. More recent works on trending topics detection [10], [11] have considered aspects of time to detect when a topic becomes popular, but otherwise do not consider the location of a topic. With the prevelance of location-based social media, such as geo-tagged tweets, the location of a social media post provides useful information about potential topics. For example, a person is more likely to post messages about sports in a stadium and food/drinks in a restaurant. In this work, we aim to address this issue by considering these spatial and temporal aspects into LDA for the purposes of topic modelling, allowing us to accurately identify specific events that are happening in a specific location or on a particular day. We further describe our proposed algorithms in Section II, and present some preliminary results in Section III.

II. Experimental Methodology

Dataset. Our dataset comprises a set of tweets collected from 01 Jan 2017 to 31 Jan 2017, using the Twitter API [12]. For this preliminary study, we focused on geo-tagged tweets that were posted within the Melbourne city area.

Data Pre-processing. Prior to using this dataset, we applied a few data cleaning steps, which includes: (i) converting all text to lower-case characters; (ii) removing standard English stopwords, such as “this”, “that”, “the”; and (iii) removing punctuation characters and numbers.

Algorithms and Baselines. Using the processed dataset, we then use our proposed algorithms and baselines to identify topics and representative keywords. The variants of our proposed algorithms and baselines are:

- **Original LDA (O-LDA).** This algorithm is the standard version of LDA as proposed in [4], which is applied on the entire dataset.
- **Author-based LDA (A-LDA).** A variation of the standard LDA where all tweets posted by the same author are aggregated as a single document [13], [5].
- **Temporal LDA (T-LDA).** Our proposed temporal version of LDA, where we extract tweets within a specific

TABLE I
SUMMARY RESULTS

		Example of Detected Topical Words	
Algo.	**Time/Space**	**Topic 1**	**Topic 2**
O-LDA	Nil	Topic O-1: humidity, wind, temperature, rising, barometer	Topic O-2: ausopen, arena, happy, australianopen, tennis
A-LDA	Nil	Topic A-1: hiring, careerarc, job, sales, opening	Topic A-2: firealarm, incident, structurefire, temperature, responding
T-LDA	01 Jan 2017	Topic T1-1: happynewyear, friends, family, beautiful, great	Topic T1-2: nye, happy, drinking, sgbrewco, fireworks
T-LDA	20 Jan 2017	Topic T2-1: bourke, street, police, closed, photo	Topic T2-2: ausopen, australianopen, tennis, rod, laver
T-LDA	26 Jan 2017	Topic T3-1: australiaday, invasionday, changethedate, fireworks	Topic T3-2: roger, federer, wawrinka, great, happy
T-LDA	29 Jan 2017	Topic T4-1: rafa, nadal, rogerfederer, federer, final, tennis	Topic T4-2: australianopen, ausopen, federer, back, great
S-LDA	Melb. Cricket Grounds	Topic S1-1: mcg, strikers, gostars, bbl, derby	Topic S1-2: starsbbl, cricket, mcg, watching, best
S-LDA	Docklands Stadium	Topic S2-1: etihadstadium, renegadesbbl, getonred, cricket, stars	Topic S2-2: docklands, melbourne, coldplay, great, night
S-LDA	Natl. Gallery of Victoria	Topic S3-1: ngvmelbourne, davidhockney, ipad, art, drawing	Topic S3-2: ngvmelbourne, viktorandrolf, exhibition, fashion, january
S-LDA	Rod Laver Arena	Topic S4-1: tennis, great, rogerfederer, rafaelnadal, rodlaverarena	Topic S4-2: australianopen, ausopen, serena, win, legend

time window, and separately run LDA on tweets of each time window. We set the time window as 24 hours based on the timestamp of tweets, which are clustered in the same group if they are posted between 00:00AM to 11:59PM of the same day.[1]

- **Spatial LDA (S-LDA).** Our proposed spatial version of LDA, where we first map tweets to specific locations based on their geo-coordinates, then run LDA on tweets associated with each location. For each location, we extract their latitude/longitude coordinates from their Wikidata entry, specifically the "coordinate location" field, and cluster tweets as belonging to a location if they are in close proximity of that location [14].[2]

Evaluation. For our initial validation, we employ a qualitative evaluation where we manually compare the keywords of the detected topics against ground-truth topics as determined from online news articles and Wikidata. For example, we identify key events on the days of 01, 20, 26 and 29 Jan 2017, as well as key activities happening within the vicinity of the landmarks of Melbourne Cricket Ground, Docklands Stadium, National Gallery Victoria and Rod Laver Arena. In future, we intend to conduct a more quantitative evaluation using the metrics of precision, recall and F1-score of detected topics and keywords against that of the ground-truth, based on a dataset with explicitly labelled ground-truth topics.

III. RESULTS AND DISCUSSION

Table I shows a summary of results obtained by our proposed T-LDA and S-LDA algorithms, in comparison to the baseline O-LDA and A-LDA algorithms. For our T-LDA algorithm, we focus on the dates of 01, 20, 26 and 29 Jan 2017, while for the S-LDA algorithm, we focus on the areas of Melbourne Cricket Ground, Docklands Stadium, National Gallery Victoria and Rod Laver Arena. The O-LDA and A-LDA algorithms represent the baselines, which are not associated with any specific time periods or spatial areas.

Results: O-LDA and A-LDA Baseline. We first examine the results of the O-LDA and A-LDA baselines. The results show that O-LDA is able to identify topics such as Weather (Topic O-1) and the Australian Open tennis tournament (Topic O-2), while A-LDA detected topics such as Careers (Topic A-1) and Fire Warnings (Topic A-2). While these are relevant topics, these are general topics that do not provide much insights into the issues covered. Next, we discuss results of our T-LDA algorithm, which considers the context of time.

Results: T-LDA Algorithm. In contrast to the baselines, all topics detected by T-LDA correspond to specific real-life events, which we manually verified from news articles and/or Wikidata entries. For example, Topic T1-1 and T1-2 are aligned to the ground truth of New Year's Day, and more specifically they can be divided into sub-topics of celebrating with family and friends (Topic T1-1) or a celebration involving drinks (Topic T1-2). The Australian Open occurred from 16 to 29 Jan 2017 and was detected as Topics T2-2 and T3-2. Although the baseline O-LDA also detected this topic (Topic O-2), our T-LDA algorithm is able to provide more details regarding these topics, such as Topic T3-2 that indicates delight at the match of Roger Federer against Stan Wawrinka, and Topics T4-1 and T4-2 that highlights the finals between Roger Federer and Rafael Nadal. In addition, we are able to detect incidents such as the Bourke Street car attack (Topic T2-1), and possible negative sentiments against the Australia Day event (Topic T3-1) as indicated by the words "InvasionDay" and "ChangeTheDate".

Results: S-LDA Algorithm. The S-LDA algorithm successfully detects topics that are relevant to the specific places they are associated with. For example, Topics S1-1 and S1-2 are both related to Cricket with the mention of cricket teams/events, such as "strikers", "starsbbl", "gostars", "bbl" and "derby". Similarly, Topic S2-1 is also related to Cricket, specifically the Melbourne Renegades cricket team that is based locally at Docklands Stadium, while Topic S2-2 is about a night concert by the Coldplay music band in the same stadium. In relation to the National Gallery of Victoria, the

[1]For this preliminary study, we adopt a simple time window of 24 hours but in future work, we plan to experiment with a varying time window that is automatically determined based on tweeting frequency and words usage.

[2]In addition, the "instance of" field gives us an indication of the type of location, e.g., "Docklands Stadium" is an instance of "multi-purpose stadium" and "cricket field", which we can use as a ground-truth for evaluating the detected topics. Subsequently in Section III, we show that the topics of Cricket and Music were detected for Docklands Stadium, coinciding with a real-life cricket match and concert.

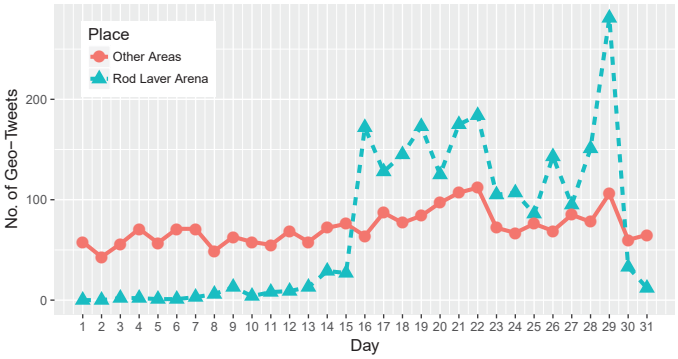

Fig. 1. An illustration of the frequency of geo-tagged tweets for Rod Laver Arena compared to other locations. Using our simple frequency-based event detection, we accurately detect the start and end of an event on 16 Jan and 29 Jan, respectively, which corresponds to the ground-truth start and end dates of the Australian Open.

Topics S3-1 and S3-2 accurately identify an art exhibit by David Hockney and a fashion exhibit by Viktor and Rolf, respectively, both held at the same location. On the same note, Topics S4-1 and S4-2 are about the Australian Open Men Finals, where Topic S4-1 accurately highlights the Mens Final involving "RogerFederer" against "RafaelNadal" at the "RodLaverArena", while Topic S4-2 talks about "Serena" Williams winning the Womens Finals.

Discussion. These preliminary results show that our proposed T-LDA and S-LDA algorithms are able to detect highly relevant and detailed topics associated with specific days and locations, in comparison with the O-LDA and A-LDA baselines that only provide an overview of more general topics. In future, we intend to further develop variations of the T-LDA and S-LDA algorithms as a joint model for detecting trending location-specific topics. For example, Figure 1 shows the tweeting frequency of geo-tagged tweets in Rod Laver Arena, in relation to an equal number of tweets randomly selected from other locations. A simple location-based trending detection mechanism is to examine the frequency of geo-tagged tweets, e.g., a trending event started if the number of geo-tagged tweets in a location doubled compared to the previous day, and similarly the event ended if the numbers halved. Using this simple but effective mechanism, we are able to detect a trending event from 16 to 29 Jan 2017 at the Rod Laver Arena, which coincides with the real-life event of Australian Open held at the same location.

IV. Conclusion

This work presents a preliminary study into incorporating spatial and temporal elements to topic modelling on Twitter, in the form of our proposed T-LDA and S-LDA algorithms. In particular, the S-LDA algorithm utilizes known landmarks listed on Wikidata to map tweets to specific landmarks based on proximity. The initial results are promising with T-LDA successfully identifying key events that are happening on specific days, and S-LDA being able to detect key activities happening in the vicinity of various landmarks.

For future work, we intend to explore the following: (i) Instead of using a fixed time window of 24 hours, automatically determine the appropriate time period to apply our topic modelling, potentially by detecting changes in frequent terms and tweeting volume; (ii) Develop a joint spatial-temporal LDA model for trending topic detection, which incorporates the location and posted time of tweets, along with key landmarks as identified from Wikidata [15]; and (iii) Extend our spatial clustering of tweets to better identify topics and activity centres that are not related to prominent landmarks, e.g., a small alley or road junction.

Acknowledgments. This research is supported by the Defence Science and Technology Group.

References

[1] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu, "Predicting flu trends using twitter data," in *Proceedings of the 2011 IEEE Conference on Computer Communications Workshops (INFOCOM'11)*, 2011, pp. 702–707.

[2] A. L. Kavanaugh, E. A. Fox, S. D. Sheetz, S. Yang, L. T. Li, D. J. Shoemaker, A. Natsev, and L. Xie, "Social media use by government: From the routine to the critical," *Government Information Quarterly*, vol. 29, no. 4, pp. 480–491, 2012.

[3] M. Conover, J. Ratkiewicz, M. R. Francisco, B. Goncalves, F. Menczer, and A. Flammini, "Political polarization on twitter," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, 2011, pp. 89–96.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[5] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proceedings of the First Workshop on Social Media Analytics (SMA'10)*, 2010, pp. 80–88.

[6] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving LDA topic models for microblogs via tweet pooling and automatic labeling," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*, 2013, pp. 889–892.

[7] S. B. Jr, G. S. Kido, and G. M. Tavares, "Artificial and natural topic detection in online social networks," *iSys - Revista Brasileira de Sistemas de Informacao*, vol. 10, no. 1, pp. 80–98, 2017.

[8] K. H. Lim, S. Karunasekera, and A. Harwood, "Clustop: A clustering-based topic modelling algorithm for twitter using word networks," in *Proceedings of the 2017 IEEE International Conference on Big Data (BigData'17)*, 2017.

[9] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *Proceedings of the 33rd European Conference on Information Retrieval (ECIR'11)*, 2011, pp. 338–349.

[10] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Göker, I. Kompatsiaris, and A. Jaimes, "Sensing trending topics in twitter," *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1268–1282, 2013.

[11] Y. Chen, H. Amiri, Z. Li, and T.-S. Chua, "Emerging topic detection for organizations from microblogs," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*, 2013, pp. 43–52.

[12] Twitter Inc., "Twitter api," Internet, 2017, https://dev.twitter.com/docs.

[13] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: Finding topic-sensitive influential twitterers," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM'10)*, 2010, pp. 261–270.

[14] K. H. Lim, J. Chan, S. Karunasekera, and C. Leckie, "Personalized itinerary recommendation with queuing time awareness," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'17)*, 2017, pp. 325–334.

[15] Wikidata, "Wikidata knowledge base," Internet, 2017, https://www.wikidata.org.