Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 72 (2023) 1161–1168

Transportation Research Procedia

www.elsevier.com/locate/procedia

Transport Research Arena (TRA) Conference

Refining agent-based travel demand models using social media data

Serra Yosmaoglu[a,*] , Diaoulé Diallo[b], Tobias Hecking[b], Alain Schengen[a]

[a]*DLR Institute of Transport Research, Rudower Chaussee 7, Berlin 12489, Germany*
[b]*DLR Institute of Software Technology, Linder Höhe, Cologne 51147, Germany*

Abstract

This work presents an approach for detecting localisable events from social media using machine learning techniques and how this information can be used in agent-based travel demand models. The approach allows for explicitly modelling irregular mobility patterns that emerge from incidences and events that gather more people as usual in certain locations, and thus, create deviating travel demands. The advantage of our approach is that it solely relies on publicly accessible social media data that can be analysed in real-time. The method is demonstrated along a case study on mapping the event-related mobility scene in Berlin (Germany).

Keywords: Machine Learning; Social Media Data; Data Mining; Location Estimation; Travel Demand; Big Data.

1. Introduction

Agent-based travel demand models represent a microscopic view on how, when, where, and why individuals move in space and have become essential for transport planning. Apart from modelling regular mobility behaviour relevant information is also needed that enables the decision makers to anticipate events and allow them to assign required resources in a study area [9]. Such information can be derived from mobility datasets (e.g. GPS data), which are considered to be expensive and time-consuming to obtain. With the gigantic growth in usage of both mobile devices and social-media in recent years, the amount of social media data including georeferences has increased dramatically and constitutes a promising and easily accessible data source for mobility modelling. In this context, Twitter has become a popular way of communicating in the recent years that people can share and exchange information regarding real-world events. Such data also enable us to detect localisable events that can be used to assess and model the mobility situation in a city. Moreover, it helps identifying and managing possible incidents and crises as well accessing mobility behaviour in pandemic scenarios.

In this paper, we first present an approach of event detection utilizing SEDTWik [14], a framework that introduces event detection in Twitter data using Wikipedia and combine it with tweet location estimation, so that localisable gatherings of people can be anticipated. The results are used to inform an agent-based model for assessing

*Corresponding author. Tel.: +49-30-67055-8150. *E-mail address:* serra.yosmaoglu@dlr.de

2352-1465
10.1016/j.trpro.2023.11.573

travel demand [10] in Berlin, Germany. The results show that the proposed method can be useful to assess how the happenings in a city can affect the mobility behaviours.

The remainder of the paper is structured into the following parts. In Section 2, we review prior studies which perform event detection and geoentity extraction from Twitter. In Section 3, we elaborate our methodology. Section 4 shows the outcomes of the event detection and agent-based travel demand model. Finally, section 5 concludes the paper and outlines possible future works.

2. Related Work

2.1. Event Detection

In the context of event detection on Twitter, events are defined as the occurrence of topics and entities of substantial volume in a certain period [4, 7].

Saeed et al. [15] differentiate between specified event detection and unspecified event detection. While for a specified event information is available and can be included in the event detection process, unspecified events use no prior knowledge and need to attend to temporal aspects of the incoming tweets. Events can be identified in Twitter streams using supervised, unsupervised and semi-supervised methods.

With a focus on specific events and associated keywords using word co-occurrences, the work of [1] is settled in the area of specified unsupervised event detection. Tweets are divided into different time windows, and a Transaction-based Rule Change Mining method is used to extract newsworthy hashtag keywords that are checked against ground-truth data from mainstream media and BBC's official website.

Zhou et al. [21] proposes the Latent Event and Category Model (LECM), a Bayesian model-based approach. This specified event detection method makes use of different linguistic features, external knowledge as well as of the meta data of a tweet. They use a table of keywords to identify relevant tweets and train a classifier that distinguishes between event and non-event tweets using features such as word frequencies, URLs and mentions. Freebase API is used to detect named entities in the tweet texts and assign semantic labels, e.g. location, that are grouped into event clusters in the final step. Since in our case events that affect peoples' mobility behaviour are not known in advance, methods for unspecified event detection are needed.

Thus, a more suitable method is SEDTWik [14], a framework for unsupervised detection of unspecified events. It can be used to identify tweet segments, i.e. hashtags, words and named entities that occur more often than expected. Wikipedia is used as external knowledge base to determine if a word is potentially relevant to an event. Since our work is based on SEDTWik, further details are provided in the methods section.

Apart from the event itself, its location is of interest as well. Therefore, Unankard et al. [19] proposed an event detection approach considering user and event locations in addition to the detection of keywords. The user location is collected from the meta data of a tweet. The event location is defined as the location term with the highest frequency, which was also recognized by a Named Entity Recognition (NER) algorithm and successfully queried in a gazetteer database. A rough approximation of the location can be achieved, however, the user location is given by the Twitter user himself and cannot be expected to be accurate. In addition, by searching a gazetteer it is not possible to find locative expressions, such as *city center*, *old town* or *town hall*.

With the help of a spatiotemporal tweet count prediction model Wei et al. [20] identify periods with deviating number of tweets that are annotated with coordinates. Using this information temporal and spatial burstiness as well as topical coherence are computed and used to infer potential event locations in time and space. A drawback of this approach is, however, that only a small fraction of tweets ($\sim$ 6%, c.f. [12]) are associated with exact coordinates, especially in the case of Germany [17].

2.2. Geo-entity Extraction from Twitter

Because the limited availability of accurately geotagged tweets, methods were developed to infer the location of tweets. This is usually done using NER models or geographic gazetteers and databases to detect mentions of locations in tweet texts [19, 20, 21]. A challenge for model-based NER is to deal with different languages since most NER models were built for English.

As Schiers et al. [16] point out, current NER models for German texts are trained with the CoNLL 2003 or the GermEval dataset [5, 8] and are not specifically designed for location detection. Therefore, they created a German dataset for NER with 15 different traffic-related classes, which was further extended [11] resulting in the only annotated German corpus in the mobility domain MobIE.

3. Methodology

As stated in the beginning, social media data (here Twitter) can be utilised to incorporate localisable events that affect mobility patterns by attracting a larger number of people into existing traffic models. This contributes to a more realistic mapping of mobility in a certain area, and thus, can for example be useful to inform traffic management strategies, location recommendation, or identify crowded places with high infection risk during pandemics. To this end, we present an extension of SEDTWik [14], an algorithm for event detection on Twitter that utilises auxiliary information from Wikipedia. SEDTWik takes a Twitter Stream from a time period as input and seeks for segments (Hashtags, user mentions, words that exist as Wikipedia page title) that are "bursty" in the sense that they occur more often in a specified time window than expected. Such bursty segments are clustered according to temporal proximity and co-occurrences in tweets, where each cluster denotes an event. Event clusters are ranked based on a so-called newsworthiness value, which is calculated by analysing segments according to their Wikipedia keyphraseness value [6]. The value indicates how often a term occurs on Wikipedia pages as anchor text compared to cases were the term occurs not as anchor text. Since in this study we focus on events taking place in Berlin (Germany), newsworthiness tables were re-calculated using a dump of the German Wikipedia instead of the English Wikipedia.

The original approach is further extended such that unspecified events are localisable by adding a self-trained geo-entity detector to the segment selection and burstiness detection. The details will be described in the following sections.

3.1. Geo-Entity Recognition

In order to include geo-entities as segments of interest in the SEDTWik event detection process we trained a NER model for place name extraction on MobIE, a German Mobility dataset [11] for training information extraction models and applied it on German tweets collected between 04/12/2021 and 15/04/2022. MobIE consists of tweets and RSS feeds which are annotated with 20 classes such as *location*, *location - stop*, *date*, *person* and *organisation*. Since we focus on Twitter data as well as the mobility domain, this dataset is particularly suitable for training a NER model to detect geo-entities mentioned in tweets.

For the development of a NER model, we used Flair [3]. Flair is a Natural Language Processing (NLP) framework which combines different types of word and document embeddings and achieves state-of-the-art results in several NLP tasks. Among multiple word embeddings, the framework offers contextualized flair embeddings [2], which consider the context of a word. This allows, for example, to distinguish between the use of a word as a name or as a location. For our work, we make use of the StackedEmbeddings class of the Flair framework. By stacking different embeddings, it is possible to combine the advantages of different word embedding models. To achieve state-of-the-art results on German text, the authors recommend to use German FastText embeddings combined with Flair embeddings. This procedure combines a traditional word embeddings method with contextualized string embeddings of Flair[1]. As training parameter, we set the learning rate to 0.1 and the mini batch size to 32.

Table 1 shows the F1-score, precision and recall of the geo-location classes of the MobIE data set. The F1-score ranges between 0.7904 for *location - stop* and 0.8926 for *location - street*. Since the dataset has just been released in the end of 2021, there are no models to compare yet. However, the high average F1-score of 0.83636 among the geo-location classes shows the model capable of identifying most geo-locations.

[1] For more information, see [14] and https://github.com/flairNLP/flair

Table 1. Flair NER model training on MobIE data set: F1-score, precision and recall

Class	F1-score	Precision	Recall
location	0.8105	0.8105	0.8105
location-city	0.8426	0.8128	0.8746
location-street	0.8961	0.9352	0.8601
location-route	0.7904	0.9124	0.6972
location-stop	0.8422	0.8660	0.8196

The advantage of using a machine learning-based model is that it learns not only to detect state, city or street names but recognizes unspecified and context dependent place mentions like “town hall”, “old town”, or “corner” as well. Through the use of contextualized Flair embeddings, the model learns where place terms are likely to be positioned in a sentence. These terms alone may not be sufficient to find associated coordinates, but placed in a context with coarse-grained geo-entities such as “prenzlauer berg”, “oranienplatz” or “brandenburg gate”[2] , these non-specific geo-terms enable exact determination of coordinates.

3.2. Localisable Event Detection

The modified version of the SEDTWik algorithm of localisable event detection consists of three steps (tweet segmentation, bursty segment extraction, and bursty segment clustering) that are described in more detail in the following:

Tweet segmentation. After removing stopwords and punctuations, SEDTWik utilizes Wikipedia to identify meaningful segments in tweet texts, i.e. unigrams or multi-grams that can be matched to the title of Wikipedia articles. In the original work [14], the authors argue that segments with more than one word are more informative than single words. Therefore, hashtags are split for uppercase letters, resulting in segments with one or multiple words. We modify the selection of segments by keeping not only segments that occur as Wikipedia articles and thus have potential to co-present an event, but also those that are recognized as geo-entities by the trained geo-entity recognizer described before and locations detected by Flair’s pre-trained NER model [3]. It was trained on the wellknown CoNLL-2003 data set and is accessible via the Flair NLP framework [3, 18]. By incorporating geo-entities, a higher number of segments and resulting segment clusters will contain locations and increase the probability of detecting localisable events. Our preliminary experiments on localisable event detection have shown that the consideration of entities referring to persons, such as ’karl lauterbauch’ or ’scholz’[3] creates noise since oftentimes politicians and celebrities are associated to specific places and then wrongly detected as an event - place combination. Consequently, in contrast to the original approach user mentions in tweets were not included as segments.

Bursty segment extraction. Subsequently, the “burstiness” score of each extracted segment is calculated by comparing the number of occurrences of each segment in a time window with the expected number of occurrences (see [14] for details) and combined with retweet and follower count as additional factors. In this work, retweet and follower count are discarded since they appeared not to be beneficial for localisable event detection in preliminary evaluations. The reason is that this amplification of tweets from Twitter users with a wider reach leads to a focus on more global and less localisable events. Instead, a geo-entity weight is introduced, which adds burstiness to each segment which contains a geo-entity.

Bursty segment clustering. The last step is grouping related bursty segments into event clusters. To do so, [14] use a variation of the Jarvis-Patrick algorithm. A network of segments is constructed, where segments are the nodes connected by an undirected edge (s_i, s_j) between segments s_i and s_j if s_i is in the list of *k_nearest_neighbors* of s j and vice versa. The *k_nearest_neighbour* lists are created by using a similarity value $sim_t(s_a, s_b)$ ([14] for detailed formula). The connected components of the resulting network denote the segment clusters representing events. In our configuration, connected components (at least two nodes) are kept while neighbourless nodes are discarded. The resulting segment clusters are ranked using the newsworthiness values of the single segments.

[2] District, square and sight, located in Berlin, Germany.
[3] German politicians.

3.3. Event implementation into an agent-based travel demand simulation model

We have chosen a microscopic, agent-based travel demand simulation called TAPAS[4] (Travel Activity PAttern Simulation) to integrate and simulate extracted events. Fig. 1 shows the model generation and is described in detail in [10]. TAPAS iteratively generates daily routines of activities for a synthetic population represented as agents. Each agent is associated with a set of sociodemographic properties like age, sex, household environment and education as well as a set of mobility options like driving license information, possession of a bicycle, public transport ticket. The daily routines are derived from empirically gathered diaries reported in MiD 2008 [13]. For each agent, TAPAS generates a daily plan consisting of a chain of trips and activities. Then for each trip a location and mode selection process is applied. A feasibility check in the matters of time and budget constraints decides whether the plan should be accepted or discarded. In the latter case a new plan is being computed. The simulation result is a list of trips for each agent with detailed information on start and end location, trip start time, travel time, activity and mode.

We assume an event that has been extracted as a surplus to an ordinary daily travel demand. This means that an agent still gets an ordinary daily routine. In addition, an event activity can be supplied if it fits into the activity pattern. We decided to constrain our model by only supplementing an event activity when it does occur after an agents last activity of a day. The decision is based on lacking empirical evidence on attending duration and classification of events. The latter would allow us to generate events that only target a subset of the agents, for example, a local event might have a bigger audience from locals, whereas a political demonstration might attract a more distributed audience.

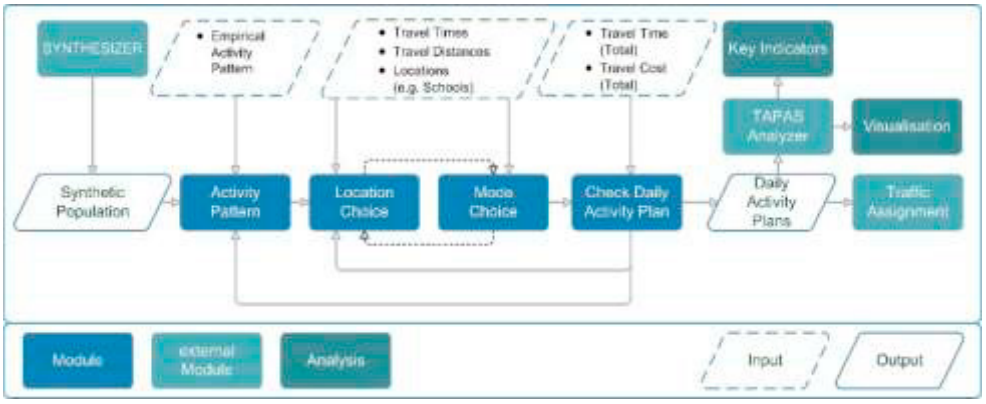

Fig. 1. TAPAS simulation flow

To demonstrate the event implementation in TAPAS, we have chosen event no. 1 (see Table 2), a football match in Berlin. It is the simplest event to implement because the size of the audience can be assumed, which reflects to the number of events supplied to agents. Additionally, the duration is straight forward and spatial allocation does not require further processing. Allocating a political demonstration with 100 thousand participants at the Brandenburg Gate location, however, is problematic because these events spread spatially.

The simulated event was distributed 22000 times, which is the capacity of the football stadium ("An der alten Försterei"). Agents were supplied with an additional event activity in their activity pattern. Since the location is known in advance, only mode choice remains necessary. The simulation output of TAPAS has been spatially aggregated to a layer of 1223 traffic analysis zones (TAZ) for Berlin.

[4] For more information, see https://github.com/DLR-VF/TAPAS

4. Results

4.1. Event detection results

We test our approach to event location detection within two periods, one between 14 to 18 February 2022 and the other between 28 February to 4 March 2022. Geo-tagged tweets from Berlin were collected via the Search API of Twitter[5]. As we want to extract geo-information directly from the tweet texts, geo-tagged tweets are not needed for the event location detection process. However, since we focus on Berlin, we use the geo-tags of the tweets to filter out all tweets that are not from Berlin. The time windows were divided into subwindows of two hours, so that one-time window with five days results in a total of 60 subwindows. In order to recognize a suitable set of geo-entities as bursty segments, a geo-entity weight of 1.2 has been found to be suitable.

Table 2. The list of events.

Event no	Occasion	Address
1	Football match	An der Wuhlheide 263
2	Antiwar demonstration	Brandenburger Tor
3	Traffic accident	Müllerstr.
4	Fire brigade operation	Emserstr.

Table 2 shows event occasions and event locations taken from the event associated tweets. In addition to a football match, we show examples of an anti-war demonstration, a car accident and a fire brigade operation.

Table 3. Events with associated segments, some geo-locations extracted from associated tweets and the total number of tweets in time window

Event no	Segments	Selection of detected geo-locations from associated tweets	Total number of tweets from time window
1	’dfb pokal’, ’pauli’, ’fcufcsp’	[...] pauli, derby, siegen, stadion [...]	6632
2	’baerbock’, ’russland’, ’münchen’, ’msc’	[...] russland, ukraine, münchen, zentrum, brandenburger, tor [...]	13046
3	[...]’schöneberg’, ’wedding’, ’straße’, ’tempelhof’, ’müllerstraße’ [...]	[...] wedding, straße, müllerstraße, friedrichshain, schöneberg [...]	13046
4	’spindlersfeld’, ’altenbraker straße’, ’hermannstraße’	[...] hermannstraße, spindlersfeld, neukölln, straße, emser, asltf [...]	13046

Table 3 shows the sets of segments that make up single events. Also, some of the geo-locations that could be recognized in the tweet texts are shown. Due to a certain error rate, there are also non-locations among the detected geo-locations. "Siegen" is a detected geo-location from event no.1, which is indeed a city in Germany, but at the same time also uses this as verb for winning. In this case, the problem is obviously the ambiguity of the term, while the segment "derby" was simply categorized incorrectly. In the case of event no. 2, "russland (russia)" and "ukraine" are correctly identified as geo-locations, but in this context the terms were subjects to the event.

[5] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets

4.2. TAPAS results

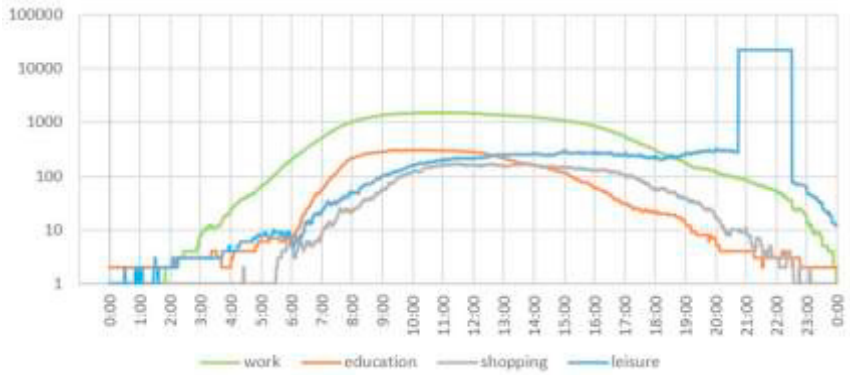

Fig. 2. Development of the number of agents in the TAZ where the event takes place

To showcase the result from TAPAS, we plotted the count of activities in the traffic analysis zone where the stadium is located (see Fig. 2). Each line marks the number of agents that are in the zone at a specific point in time for a specific activity throughout the day. One must note that the y-axis is in logarithmic scale. Four main activities, work, education, shopping and leisure are visualised, where the event activity is a sub-group of the leisure activity. One can see a sudden rise of agents from 275 to 22281 in the zone at around 20:45 and a sudden drop at around 22:30 from 22094 down to 77 agents for the leisure activity which is about the time slot of the football match. This result was to be expected since the event was supplied 22000 times to agents as a special subgroup of the leisure activity.

5. Conclusion

In this paper, we have showcased the potential of evaluating social media data to parameterise travel demand models. So far, agent-based travel demand models usually map daily routines such as work, education and shopping but are incapable of modelling temporally changing travel patterns caused by serendipities and events. By integrating machine learning models for event and location detection in Twitter data, it is possible to detect gatherings of people that lead to mobility behaviour that deviates from “normal”. The proposed method can thus be used for more realistic travel demand modelling as well as event related traffic management and planning. It has also relevance in crisis situations or in the context of pandemics where it is crucial to adequately assess in realtime how people move and gather in a certain area and to detect locations that are occupied more frequently than usual.

In future works, our method should be evaluated also for other locations than Berlin. It is especially of interest whether there are differences in urban and rural regions for which the population density and the coverage of associated social media data differ. More research is also needed on the methodological level to model mobility behaviour from social media data on different levels of granularity (city-level, country-level, etc.). Additional research is required in the context of event classification, like size or target audience distribution, in order to supply more specific agents with events. To this end, the geographical knowledge bases could be used in the location detection step to aggregate mentions of specific places into broader geographic regions.

Acknowledgements

Authors Serra Yosmaoglu, Diaoulé Diallo, Tobias Hecking and Alain Schengen were funded by the German Federal Ministry for Digital and Transport under grant agreement FKZ19F2211A.

References

[1] Adedoyin-Olowe, M., Gaber, M. M., Dancausa, C., Stahl, F., Gomes, J. B., 2016. A rule dynamics approach to event detection in Twitter with its application to sports and politics. Expert Systems with Applications, pp.351-360.

[2] Akbik, A., Blythe, D., Vollgraf, R., 2018.MobIE: Contextual String Embeddings for Sequence Labeling. COLING 2018, 27th International Conference on Computational Linguistics, pp. 1638-1649.

[3] Akbik, A., Bergmann, T. , Blythe, D., Rasul, K., Schweter, S., Vollgraf, R., 2019. MobIE:FLAIR: An easy-to-use framework for state-of-the art NLP. NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pp. 54-59.

[4] Becker, H., Naaman, M., Graano, L., 2011. Beyond trending topics: Real-world event identification on twitter. Proceedings of the International AAAI Conference on Web and Social Media, Vol. 5., No. 1.

[5] Benikova, D., Biemann, C., Reznicek, M., 2019. NoSta-D Named Entity Annotation for German: Guide- lines and Dataset. In Proceedings of the Ninth International Conference on Language Resources and Evaluation, pp.2524-2531.

[6] Chenliang, L., Sun, A., Datta, A., 2012. Twevent: Segment-based Event Detection from tweets. Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 155-164.

[7] Dou, W., Wang, X., Ribarsky, W., Zhou, M., 2012. Event detection in social media data. IEEE VisWeek workshop on interactive visual text analytics-task driven analytics of social media content, pp. 971-980.

[8] Faruqui, M., Padó, S., 2010. Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In KONVENS, pp.129-133.

[9] Fernández Vilas, A, Díaz Redondo, RP, Ben Khalifa, M. Analysis of crowds' movement using Twitter. Computational Intelligence. 2019; 35: 448– 472. https://doi.org/10.1111/coin.12205

[10] Heinrichs M., Krajzewicz D., Cyganski R., von Schmidt, A. (2017) Introduction of car sharing into existing car fleets in microscopic travel demand modelling. Personal and Ubiquitous Computing, 1-11. Springer. doi: 10.1007/s00779-017-1031-3. ISSN 1617-4909

[11] Hennig, L., Truong, P., Gabryszak, A., 2021. MobIE: A German Dataset for Named Entity Recognition, Entity Linking and Relation Extraction in the Mobility Domain. Proceedings of the 17th Conference on Natural Language Processing, pp. 223-227.

[12] Kruspe, A., Häberle, M., Hoffmann, E. J., Rode-Hasinger, S., Abdulahhad, K., Zhu, X. X., 2021. Changes in Twitter geolocations: Insights and suggestions for future usage. arXiv preprint arXiv:2108.12251

[13] Lenz B, Nobis C, Köhler K, Mehlin M, Follmer R, Gruschwitz D, Jesske B, Quandt S. (2010) Mobilität in Deutschland 2008. DLR Project Report, http://daten.clearingstelle-verkehr.de/223/ http://elib.dlr.de/68010/. Accessed 13 May 2022

[14] Morabia, K., Murthy, N., Malapati, A., Samant, S., 2019. SEDTWik: Segmentation-based Event Detection from tweets Using Wikipedia, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, pp. 77-85.

[15] Saeed, Z., Abbasi, R. A., Maqbool, O., Sadaf, A., Razzak, I., Daud, A. et al., 2019. What's happening around the world? a survey and framework on event detection techniques on twitter. Journal of Grid Computing, 17(2), pp. 279-312.

[16] Schiersch, M., Mironova, V., Schmitt, M., Thomas, P., Gabryszak, A., Hennig, L., 2018. A German corpus for fine-grained named entity recognition and relation extraction of traffic and industry events. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation.

[17] Sloan, L., Morgan, J., 2015. Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. PloS one, 10(11), e0142209.

[18] Tjong Kim Sang, E. De Meulder, F., 2019. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pp. 142-147.

[19] Unankard, S., Li, X., Sharaf, M. A., 2015.Emerging event detection in social networks with location sensitivity. World Wide Web, pp. 1393 1417.

[20] Wei, H., Zhou, H., Sankaranarayanan, J., Sengupta, S., Samet, H., 2019. Delle: detecting latest local events from geotagged tweets. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Analytics for Local Events and News, pp. 1-10.

[21] Zhou, D., Chen, L., He, Y., 2015. An unsupervised framework of exploring events on twitter: Filtering, extraction and categorization. In Proceedings of the AAAI conference on artificial intelligence, pp. 2468–2475.