JOSIS

JOURNAL OF SPATIAL INFORMATION SCIENCE
Number 24 (2022), pp. 31–61
doi:10.5311/JOSIS.2022.24.167

RESEARCH ARTICLE

Extracting locations from sport and exercise-related social media messages using a neural network-based bilingual toponym recognition model

Pengyuan Liu[1], Sonja Koivisto[1], Tuomo Hiippala[2], Charlotte van der Lijn[1], Tuomas Väisänen[1], Marisofia Nurmi[1], Tuuli Toivonen[1], Kirsi Vehkakoski[3], Janne Pyykönen[3], Ilkka Virmasalo[3], Mikko Simula[3], Elina Hasanen[3], Anna-Katriina Salmikangas[3], and Petteri Muukkonen[1*]

[1]Department of Geosciences and Geography, University of Helsinki, Finland
[2]Department of Languages, University of Helsinki, Finland
[3]The Faculty of Sport and Health Sciences, University of Jyväskylä, Finland

Received: May 31, 2021; returned: September 5, 2021; revised: September 29, 2021; accepted: December 24, 2021.

Abstract: Sport and exercise contribute to health and well-being in cities. While previous research has mainly focused on activities at specific locations such as sport facilities, "informal sport" that occur at arbitrary locations across the city have been largely neglected. Such activities are more challenging to observe, but this challenge may be addressed using data collected from social media platforms, because social media users regularly generate content related to sports and exercise at given locations. This allows studying all sport, including those "informal sport" which are at arbitrary locations, to better understand sports and exercise-related activities in cities. However, user-generated geographical information available on social media platforms is becoming scarcer and coarser. This places increased emphasis on extracting location information from free-form text content on social media, which is complicated by multilingualism and informal language. To support this effort, this article presents an end-to-end deep learning-based bilingual toponym recognition model for extracting location information from social media content related to sports and exercise. We show that our approach outperforms five state-of-the-art deep learning and machine learning models. We further demonstrate how our model can be deployed in a geoparsing framework to support city planners in promoting healthy and active lifestyles.

Keywords: digital geography, deep learning, geoparsing, georeferencing, social media, sports geography, toponym recognition

1 Introduction

Cities are the physical concentrations that facilitate interaction between people and things (physical activity environment, e.g., built environment and public green space). Therefore, understanding the human dimension of cities and how people experience the physical environment is a crucial aspect to understand urban places and urban geography [45]. It has been widely seen that among all the activities occurring every day in the cities, the geographies of sport and exercise-related activities are largely neglected in the existing literature [63, 104].

As an inseparable part of society, sport and exercise-related activities are inherently connected to the discipline of geography, in particular, concerning time, space, communities, mobilities, and identities [6, 18, 104]. Therefore, how we understand the sport and exercise-related activities in different places aligns with the continued transformation of space and attitudes towards sport [104]. The interactions between cities' inhabitants and physical open spaces and sport facilities provide insights for geographers to study how places are perceived and represented with enriched sporting activities and understand the underlying socio-economic characteristics of the urban areas [55, 83].

Sport activities in urban spaces can take a variety of forms, ranging from formal participation in competitive club sports or organised exercises such as basketball matches, to more casual engagements (or "informal sports" [5]) in active leisure such as jogging and cycling to work, or lifestyle sports like skateboarding on the supermarket car park, or outdoor play like rope skipping between buildings. Existing research that attempts to understand socio-ecological relationships between human sporting activities and urban spaces commonly focuses on formal participation or organised sports that use certain facilities [10, 60, 86] or comparably more casual physical activities in public green spaces [24, 71, 83]. One of the significant advantages of analysing registered sports facilities and green spaces is because they often have officially published statistics (e.g., *Statistics Finland*[1]) that eases the difficulties in data collection. Even without published data, they are more straightforward to collect data and observe from specific known locations. However, the impact of "informal sport" activities and unbuilt physical activity environments has been largely neglected in scientific contexts. The locations of those sporting activities are more arbitrary and difficult to observe.

In the recent years, social media is increasingly recognised as a valuable source of user-generated geographical information on physical activities and sport in the urban environment [105]. The desire of self-presentations [108] of users to their friends or even strangers has promoted the use of social media (such as Twitter or Instagram) into major platforms from which to communicate and exchange information regarding a wide variety of topics, including also recreational sporting activities. The integration of social media platforms with highly mobile devices such as smartphones opens up opportunities for users to upload their sport and exercise-related content anywhere they prefer with accessible network connections. Thus, it allows researchers to extensively investigate all sport simultaneously, including those "informal sports" activities and unbuilt environments at more arbitrary locations. Despite existing geographical studies having identified that sporting activities carried out on social media (most are Twitter) have connections to the use of urban space [44, 66, 83], the data they used are often tweets with precisely geolocated coordinates. However, our understanding of the role played by social media in the social

[1] https://www.stat.fi/index_en.html

construction of the place has been limited by the fact that only a small percentage of social media posts are precisely geolocated (e.g., according to Sloan and Morgan [89], 0.85% of tweets are geolocated). In June 2019, Twitter further decided to remove the ability of users to add precise location information (geo-coordinates pairs) in their text-based posts[2], and such a change presumably will reduce the number of geolocated tweets [50]. Although Twitter still allows users to geotag their content using *Twitter Place*[3], it has raised a significant challenge for researchers to locate tweets precisely and study the connections between places and online content.

One solution to the problem mentioned above is to investigate geographical information available in a social media content. This is an active area of research known as *geoparsing*. Geoparsing is an algorithmic toponym resolution process of converting text-based descriptions of places names into their corresponding spatial coordinates [34,36,42,80]. The process of geoparsing is usually separated into two consecutive steps: *toponym recognition* and *toponym resolution*. The first step recognises toponyms (i.e., location-indicative words) from text, and the second step assigns a location mentioned in the text to a pair of suitable geographic coordinates and sorts out any possible place name ambiguity. This paper focuses on the first step, namely toponym recognition, which can be conceptualised as a part of a broader task known as Named-Entity Recognition (NER, a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organisations, locations, etc.)

Twitter and other social media platforms are a rich source of information on real-life language use [46]. Thanks to the increasing language diversity on the Internet, social media platforms are multilingual [33]. Multilingual users pose a challenge to toponym recognition, as previous research has mainly focused on recognising toponyms in a single language and the English language in particular [4,58,59,101]. When collecting spatial data based on the toponyms in the text in a bilingual or multilingual setting, it is often needed to apply language-specific models on different languages separately, which consequently complicates the data collection process. Meanwhile, language-specific models struggle to handle "code-switching" in social media where two or more languages are often presented in one piece of text. Moreover, social media often face the challenge of various language irregularities such as informal sentence structures, inconsistent upper and lower case, dialects, name abbreviations, and misspellings [37, 49, 101, 102]. Such a high degree of variation in the *form* of text presents a challenge for Natural Language Processing (NLP) models [14]. Therefore, to extract information about location-specific activities such as sport and exercise from social media data, one must overcome the challenges of multilingualism and linguistic variation.

Our study in this paper focuses on Finland because it is well acknowledged as a successful "sports nation" [64]. Therefore, Finland offers an excellent site for the analysis of how sport contributes to space and place meaning within the displine of Digital Geography. Hiippala et al. [47] revealed that four out of five Twitter users with a predicted home location in Finland use multiple languages on the platform. In most cases, the languages used are Finnish and English, but the distribution of these languages is not balanced among users or locations. Despite that, some research has led to the development of NER tools that can be used for the toponym recognition tasks for Finnish language [96, 99] by keeping location-relevant information as the only output. There are remaining concerns

[2]https://twitter.com/TwitterSupport/status/1141039841993355264

[3]https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/place

regarding the performance of those off-the-shelf NER tools in processing user-generated text. Resources for training NER models are often constrained to narrow domains, which have linguistic conventions different from those used in social media.

In this paper, we propose a neural network-based bilingual (English and Finnish) toponym recognition model (TRM) to extract locations from social media. TRM extends a general recurrent neural network model and is combined with the state-of-the-art Transformer-based deep learning language model *BERT* [30] to create contextual embeddings for natural language understanding, together with the word- and character-level embeddings for the toponym recognition task in social media messages.

In this paper, we show that

- pre-trained multilingual language models can be fine-tuned to recognise toponyms in multiple languages with relatively little training data, leading to greater coverage of social media content;
- TRM can be adopted to extract locations from sports and exercise-related tweets, and we explore the potential of social media as a source of information on sports and exercise in cities based on the aggregated spatial data.

2 Background

2.1 Urban places and user-generated content

Analysing how places are perceived and represented is crucial to interpret the underlying social and spatial practices involved with enriched human activities such as political, social and economic activities in space. The analysis of human conceptualisations of the space often involve categorisations of some kind. Such summarisation and categorisation processes of representative geographical phenomena inform us of the understanding of socio-spatial practices in the places of a given space. Thus, understanding the representation of place is a central problem in geographical studies, and such representations of places have a strong connection with information science and information systems [81].

Within the discipline of Digital Geography, place representation often refers to the overall information available in a target geographic area for a given data source [7, 8]. The spatial and social structures of local communities in a city lead to certain collective human activities patterns [91]. With the concept that users on digital platforms are sensors of places [39], user-generated content (UGC) can help to "sense" this type of information from urban environments, focusing on the interactions between users and neighbourhood infrastructures. Thus, UGC provides a unique insight to places with abundant information of sentiment as well as relationships between individuals, groups, and the physical environment [84]. GIScience research thereby focuses on how corresponding spatio-temporal patterns of the content production from various platforms and heterogeneous data streams can be explored, extracted, validated and aggregated. In turn, such information enables us to analyse everyday spatial processes and to gain knowledge about places, especially with respect to collective human dynamics [91].

Due to the potential of social media platforms for exploring human activities in space and the narrative of places [1], social media platforms in general, and Twitter in particular, have been at the centre of data-driven analysis in GIScience and quantitative geography for about a decade [73]. Location-based social media data are used widely in

urban research ranging from urban form and structure to everyday activity practices of people [26, 51, 52, 68]. Being one of the most important activities that is happening in cities every day [62, 66, 107], sport and social media have presented a longstanding mutually beneficial relationship with each other [22]. The use of social media for sport and exercise-related purposes prompts sporting activities in the physical environment and vice versa [12]. Therefore, sport and exercise-related activities carried out on social media platforms can provide rich information regarding places, the use of space, and people's experiences of landscape [15, 43, 44, 98]. However, there is a lack of in-depth quantitative studies of sport and exercise-related activities in the literature using social media messages as a source of data. Therefore, one of the major novelties of this paper, as mentioned in the *Introduction*, is to explore the potential of social media, with Twitter in particular, as a source of information on sport and exercise in cities based on the collected spatial data. For the scope of this paper, we focus on the analysis of sport and exercise-related activities in Helsinki, which is the capital city of Finland.

2.2 Sport and exercise-related activities in Finland

Sport and exercise-related activities have played an integral role in the social life of Finland's population. The country has one of the highest sport participation rates in the whole of Europe. Studies have emphasised that for the population over 15 years old, the percentage of those practising sport at least once a week in Finland is 76% and only 4% population declared that they never do exercise or participate in sports [35]. Therefore, Finland is well acknowledged as a successful "sports nation" [64]. As the most populous region of Finland, Helsinki Metropolitan Area (HMA, including the central cities of Helsinki, Vantaa, Espoo, and Kauniainen) is the only metropolis in the country which attracts over 1 million population to settle in [90]. Among the cities in HMA, their urban planning strategies highlight the promotion of physical activity as a spearhead project (e.g., city of Helsinki's Physical Activity Programme [20], city of Espoo's Exercise Classes [19]), recognising that physical inactivity is one of the most significant factors contributing to the deterioration of wellbeing. Although Finnish people are generally active in physical activities and Finnish law requires equal opportunities for all citizens to access sports facilities [94], evidence shows that the participation and the attitudes towards sports are segregated and decided by inhabitants' social status (e.g., education background, incoming level) [55, 75, 85].

Due to an overall high willingness of the population to participate in sport and exercise-related activities, Finland offers an excellent site for analysing how to understand urban places with sporting activities. Being one of the most popular social media platforms in Finland [21], Twitter provides a crucial opportunity for researchers to investigate the information geographies of sporting activities using the social media content produced online. However, due to the growing awareness of privacy concerns, locating and collecting those recreational sporting activities from Twitter is increasingly difficult. As mentioned in the *Introduction*, Twitter has announced the removal of its precise geo-tagging feature. Such a change of Twitter's policy prompts methodological developments to recognise and geolocate tweets [50].

2.3 Locating sport and physical activities: toponym recognition tools

One of the primary approaches is to identify the candidate locative references in the text, and such a task is defined as *geoparsing* [80]. As mentioned in the *Introduction*, the process of geoparsing is usually separated into two consecutive steps: *toponym recognition* and *toponym resolution*. Many existing geoparsing research has heavily relied on off-the-shelf NER tools for the toponym recognition step [28,41,58,59] because toponym recognition is often a sub-task of NER by keeping only locations as the output. However, research has identified that the performance of many existing NER tools (e.g., Stanford NER [72]) is limited when conducted on informal text from user-generated content [49, 101, 102]. To address such a limitation and improve toponym recognition from social media messages, Wang et al. [101] introduce a Neuro-net ToPonym Recognition (NeuroTPR) model targeting the language irregularities associated with social media text to recognise locations. Their proposed model has several designed features on top of a general bidirectional recurrent neural network to address the task of location recognition in social media messages. The model achieves the state-of-the-art performance tested on GeoCorpora [100] and their proposed Twitter dataset, which shows a significant technological advance addressing linguistics variations in social media text compared to most off-the-shelf name entity recognition tools such as Stanford NER.

However, most tools in geoparsing are developed for English. This is partially due to that as a high-resourced language, English has the most abundant and well-documented datasets and information. Even though much effort has been devoted to developing multilingual NER tools for both high- and low-resourced languages, few research projects have focused on developing multilingual geoparsing tools [27]. This has significantly limited the data collection process for social media platforms, especially for Twitter which has an increasing language diversity on the Internet [47]. It is often needed to apply language-specific models (i.e., toponym recognition models or NER tools) on different languages separately to cover a broader language use on Twitter, which complicate the data collection process. Multilingual or bilingual language models have the potential to address such a limitation and ease the data collection process by applying to different languages directly. Chen et al. [16] developed a multilingual geoparsing workflow based on machine translation, consisting of three major steps. Firstly, a machine translating tool will translate different languages into English, and then based on Condition Random Fields [37], a trained English geoparser will find locations in the translated text. Finally, their multilingual geoparser uses the word alignment information to match the locations identified with those in English and original languages and converts identified locations into geo-coordinates. The framework has been tested with Chinese, Arabic and English, and its performance based on their tests has been robust. Even though the framework remains the ability, to extend to other languages, in an English-Finnish context, for example, such a tool is heavily dependent on translation quality, which is a major concern when applying this workflow to social media studies.

To the best of our knowledge, no end-to-end bi- or multilingual geoparsing tool that can be directly used in an English-Finnish context has been developed, especially for toponym recognition on social media. The primary contribution of our paper is to propose a bilingual toponym recognition model that can extract locations from social media messages (i.e., tweets).

3 Methodology

3.1 Model architecture

The proposed TRM model is based on the BiLSTM–conditional random field (BiLSTM-CRF) model proposed by Lample et al. [65]. This is widely considered as a classical framework for general NER tasks [101]. With this framework, we include several improvements inspired by Wang et al. [101] to develop our TRM model, as shown in Figure 1.

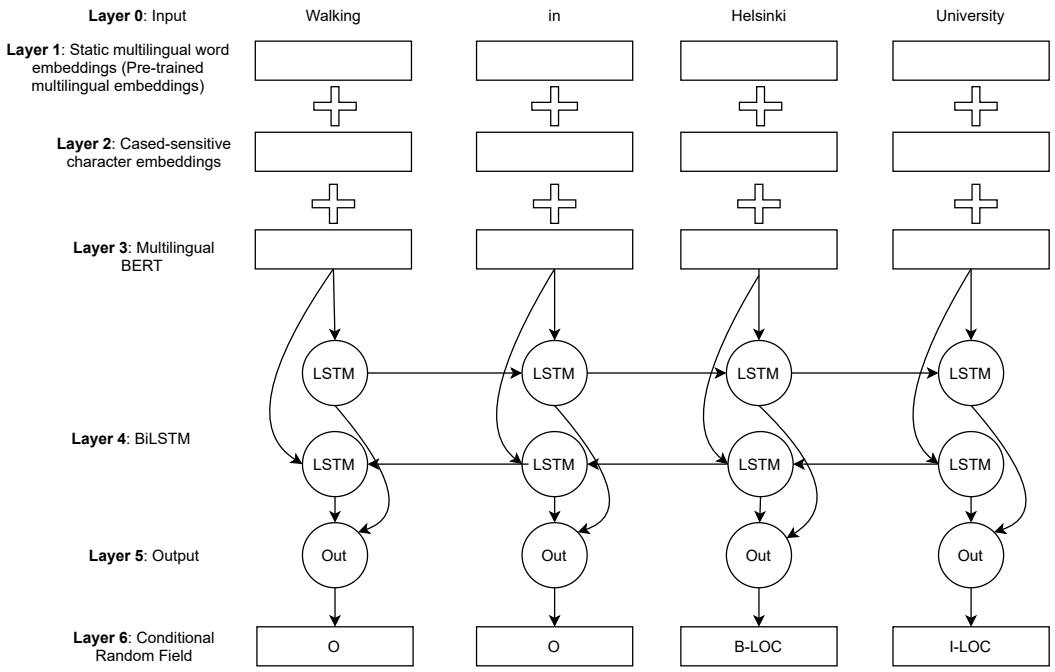


Figure 1: Overall framework for TRM with an English text as the example.

We present TRM from top to bottom using the example in Figure 1, characterising the layers of the proposed neural network. Layer 0 takes each individual word of a tweet as the input to the model. Then, each word is represented as vectors using three different approaches in the next three layers. Layer 1 adopts static multilingual word embeddings from Facebook named *MUSE* (Multilingual Unsupervised and Supervised Embeddings) [23] which has been successfully used for a wide range tasks (e.g., multilingual speech detection [11], sentiment analysis [88], informational retrieval [79], etc.). In terms of "static", vector representation for a given word in the model's vocabulary remains the same regardless of the context in which the word appears. MUSE attempts to learn a shared embedding space for multiple languages, in which the vectors for words with similar meanings across different languages are close to each other in the embedding space. Taking English word "cat" and Finnish word "kissa" (cat) as an example; traditionally, if we separately train two word embedding models in two languages, although both "cat" and "kissa" refer to the same kind of animal, the cosine distance between these two words could be far from each other as they are in the different vector spaces. MUSE will align two embeddings into the

same vector space; therefore, the words "cat" and "kissa" can have similar or even the same vectors. The published multilingual word embeddings[4] have a coverage of 30 languages (including Finnish) aligned in a single vector space. The use of word embeddings helps the model to learn whether a word refers to a location according to the specific context where the word is used.

Character embedding is adopted as Layer 2 to model each word as a sequence of characters. The character embeddings are modelled by using BiLSTM architecture [25]. As mentioned by Wang et al. [101], character embeddings are good at handling the high linguistic diversity and variations of the user-generated text. The use of this layer can aid the framework to still be able to capture the semantic meanings of a word when a user misspelling in their text (e.g., miss typing the word "location" into "locatuon", or missing any character in the word completely).

Layer 3, as shown in Figure 1, is a multilingual BERT used to capture the different semantics of a word under varied contexts. BERT is a Transformer-based deep learning technique for natural language processing (NLP) published by Google [30]. It has achieved state-of-the-art results in many natural language processing tasks. In contrast to the static word embeddings provided in Layer 1, BERT provides a dynamic (contextualised) word embedding for a word by modelling the context (i.e., sentences) where the word is used. For example, for two sentences of "I went to the river *bank*. I went to the *bank* to make a deposit", traditional word embeddings will generate one embedding for the word "bank"; however, BERT is able to capture the context of the sentence and generate different embeddings for the word "bank" under different use contexts. The pre-trained multilingual BERT adopted in this paper was published by Google that covers 104 languages[5].

The three layers mentioned above model a word into three individual representation vectors, and those representation vectors are then combined to represent each input word with a large vector. As shown in Figure 1, these vectors are then used as the input to Layer 4, which is a Bidirectional LSTM (BiLSTM) layer that consists of two LSTMs taking the input in a forward and backward direction, respectively. As such, BiLSTMs effectively enlarge the information available to the network (e.g., knowing the immediate following and preceding information for a target word). Layer 5 is a fully connected layer that combines the two LSTM layers' outputs. Layer 6 is a CRF layer that performs sequence labelling on the output of Layer 5. The CRF layer uses the standard IOB model, the same as in Wang et al. [101] to label each word but focuses on locations. Thus, the framework annotates each word with tags "B-LOC" (i.e., the beginning of a phrase which refers to a location), "I-LOC" (i.e., inside a phrase which refers to a location), or "O" (i.e., outside a phrase which refers to a location).

Our model is a variation and improvement of NeuroTPR [101] for bilingual toponym recognition tasks. Compared with NeuroTPR, we removed the caseless character embedding layer to simplify the model architecture, and we replaced ELMO [53] with BERT because BERT has demonstrated stronger capabilities in many NLP tasks [30]. We further remove the layer of part-of-speech tagging layer to reduce the impact from distinctive grammatical features of English and Finnish during the training. See Section 4.2 for further discussion. Although our model is designed for bilingual geoparsing tasks, the use of pre-trained multilingual models in our framework: MUSE and multilingual BERT are a sig-

[4]https://github.com/facebookresearch/MUSE
[5]https://github.com/google-research/bert/blob/master/multilingual.md

nificant advantage. That is, the TRM architecture can be easily adapted to other language pairs.

3.2 Data preparation

3.2.1 Training data

Wang et al. [101] demonstrates that other domains of information sources (i.e., Wikipedia articles) can also be useful training data for deep learning models to perform toponym recognition on social media text. Therefore, their paper indicates that deep learning models may have the generalisability to learn across the domains. Also, a joint learning process based on a combination of social media text and Wikipedia articles as training data can further benefit the model to identify correct toponyms. Therefore, in this paper, as a model designed for bilingual geoparsing tasks, we trained our TRM on a combined Finnish and English dataset following a similar process of data preparation introduced by Wang et al. [101] in their paper.

We adopted an automated annotation workflow proposed in Wang et al. [101] for English training data preparation. This workflow operates the first few paragraphs of English Wikipedia articles that often annonate the entities mentioned in the text using hyperlinks. An annotated training data set is generated by extracting these paragraphs from a Wikipedia dump. With the help of Infobox[6], only the phrases whose hyperlinks linking to articles about geographic location are possessed. Since the data are generated for training our TRM in toponym recognition for tweets, we manufactured our data to share a more consistent form as tweets by splitting the Wikipedia paragraphs into sentences and retaining only those within 280 characters (the maximum allowed by Twitter[7]). The sentences extracted from Wikipedia are automatically labelled with "B-LOC" (i.e., the beginning of a phrase which refers to a location), "I-LOC" (i.e., inside a phrase which refers to a location), or "O" (i.e., outside a phrase which refers to a location) as already mentioned in the previous section in the CoNLL2003 format [95]. We prepared and annotated 3000 phrases from English Wikipedia with the help of this workflow. In addition to the sentences from English Wikipedia articles, we followed the same concept of data preparation described in Wang et al. [101] and obtained 300 English tweets from a published NER benchmarking Twitter dataset *WNUT 2017 Shared Task on Novel and Emerging Entity Recognition* (WNUT 2017) [29]. WNUT 2017 is a dataset that contains real tweets annotated by human annotators, and the tweets contain toponyms along with other types of entities. The 300 tweets were selected by filtering out tweets that contain toponyms, and we kept only toponyms in the annotations as a final English Twitter dataset used for training TRM.

Compared with English, Finnish data preparation is less straightforward, considering it is a low-resourced language. The workflow [101] for extracting and annotating sentences from English Wikipedia articles cannot be applied for Finnish Wikipedia articles because Finnish articles have fewer hyperlinks. The Infobox of the articles commonly lacks information regarding geographical entities. Therefore, we adopted the Turku NER corpus from TurkuNLP [70] to prepare our Finnish dataset. The Turku NER corpus consists of a range of text domains, including news, user-generated text such as blog posts, and legal texts for NER tasks, and the corpus annotation marks mentions of the person (PER), organisation

[6]https://en.wikipedia.org/wiki/Wikipedia:List_of_infoboxes/Place
[7]https://developer.twitter.com/en/docs/counting-characters

(ORG), location (LOC), product (PRO) and event (EVENT) names as well as dates (DATE). At first, we keep sentences with both ORG and LOC as there are commonly ambiguities between these two categories. For example, as mentioned in the Turku NER corpus annotation guidelines[8], buildings, facilities and similar entities referred to by the name of an organisation are annotated as ORG, such as "Stockmann" (a Finnish retail company) in "menin Stockmannille" (I went to Stockmann) is marked as ORG. Still, we might consider it as a toponym to be identified as a location if our targeted studying area is the city of Helsinki (see Section 5). Therefore, we manually examined each sentence and solved such ambiguities and further filtered out sentences that only have toponyms indicating locations. In general, we change and annotate the following to LOC:

- administrative place names, such as neighbourhoods, towns, cities, states, and countries;
- names of natural features, such as rivers, mountains, and beaches;
- names of facilities and landmarks, such as roads, train stations, buildings, bus stops and airports.

Such a principle of annotation will be further applied to annotate the Twitter datasets we collected (see next subsection). The final Finnish dataset consists of 1587 sentences.

3.2.2 Test data

The original Twitter dataset contains 38,487,766 tweets from Finland and Estonia covering the time frame between 08.09.2006 and 15.04.2020, and it was collected and cleaned up by the *Digital Geography Lab* at the University of Helsinki [47]. The dataset contains 23,248,531 Finnish tweets and 12,042,826 English tweets. Hypothetically, if all tweets mention sports and locations, a bilingual model can collect 34% more data than an English-only language model and 66% more data than a Finnish-only language model. Therefore, it is necessary to consider bilingualism when analysing social media data in Finland. As shown in Figure 2, we designed the workflow to prepare the test data extracted from the original Twitter dataset. Firstly, we extract 200,000 tweets that contain certain sports-indicative words for English and Finnish tweets, respectively. The language is decided by using the attached language code supported by Twitter in the JSON file we collected. The sports-indicative words are decided by a list of words containing nouns and verbs as shown in Table 1.

We are aware that some of the sports may be missing in the word list and such a matching-on-form approach is relatively naive to filter out sport and exercise-related tweets, the corresponding tweets contain a large amount of "noise". For example, tweets similar to "we keep the computer running" would also be identified as sport and exercise-related tweets in the dataset. However, advanced methodological development of sporting activities classification is beyond the scope of this paper. Further discussions will be provided in Sections 5 and 6.

After the first step, we randomly sampled 9000 tweets in English and 9000 tweets in Finnish, and we then manually examined the data, selected and annotated 750 sport and exercise-related tweets with toponyms to be recognised following the guidelines mentioned in Section 3.2.1 in English as well as in Finnish. Therefore, the final dataset consists of 1500 labelled tweets, and it would be used as the test data to evaluate the performance of our proposed TRM.

[8]https://github.com/TurkuNLP/turku-ner-corpus/blob/master/docs/Turku-NER-guidelines-v1.pdf

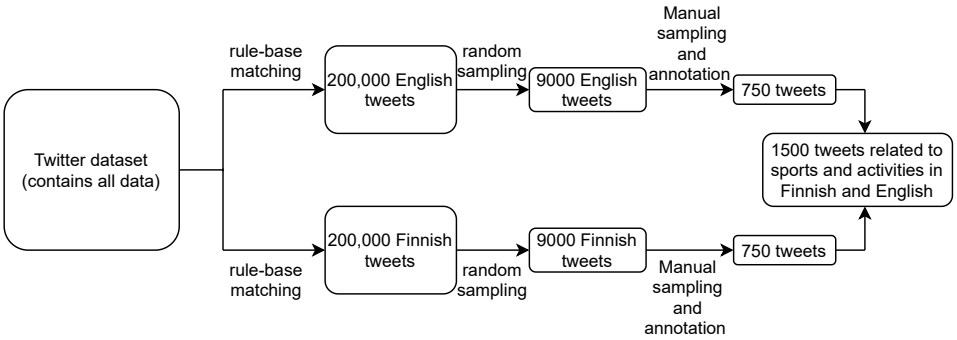


Figure 2: Test data preparation.

Table 1: List of sports-indicative words in English and Finnish.

English	Finnish
walk, walking	kävely, kävellä, käveleminen
running, run	juoksu, juosta, juokseminen
jog, jogging	lenkki, lenkkeily, lenkkeillä
hike, hiking	patikointi, patikoida, patikoiminen
trek, trekking, bicycle, bike, biking, cycling	pyörä, pyöräily, pyöräillä, pyöräileminen
exercise, exercising, workout, training, sport, sporting	treeni, urheilu, liikunta, treenata, treenaaminen, urheilla
gym	kuntosali
sweat, sweating	hiki, hikoilla
ski, skiing	hiihto, hiihtää, hiihtäminen
skate, skating	luistella, luisteleminen, luistelu
ice-hockey, hockey	jääkiekko, lätkä,
basketball	koripallo, koris
football, footy	jalkapallo, futis
tennis	tennis
badminton	sulkapallo
floorball	sähly, salibandy
volleyball	lentopallo, lentis
beachvolley	rantalentopallo
dance, dancing	tanssi, tanssia, tanssiminen
yoga	jooga
swim, swimming	uinti, uida, uiminen
kayak, kayaking, rowing, canoe, canoeing, sail, sailing	meloa, melonta, soutaa, soutaminen, kajakki, kanootti, pujehtia, purjehdus

In both training data and test data, the level of granularity of the geographical areas mentioned in the text varies from the geographical scales of countries to point-of-interest. That is, the toponyms for TRM to recognise can be the names of countries (e.g., “Finland”), cities (e.g., “Helsinki”) or more fine-grained toponyms such as streets and buildings (e.g., “Stockmann”). Although TRM in this paper is primarily used to recognise toponyms at the

fine-grained resolution in the city of Helsinki (see Section 5), our model can be adopted to identify location-indicative words at any geographical granularity.

4 Model training and experiments

We implemented our TRM in Python using the FlairNLP platform [3] with Pytorch [77] as the backend. We trained our model on Google Colab[9], which provides powerful Nvidia GPU supports[10]. The training starts with an initial learning rate of 0.1 using Adam [61] with a loss function of Cross-Entropy Loss. The learning rate will decrease by half if the model performance during the training is not improved for three consecutive iterations until the learning rate is too low (learning rate < 0.000195). The source code used for this paper is available on GitHub[11].

The evaluation metrics used in the experiments are Precision, Recall, and F-score, which have been widely used in previous studies, such as [57, 67, 101]. Precision measures the fraction of correctly identified toponyms (true positives) among all toponyms (including true positives and false positives – falsely recognised toponyms) annotated by a model. Recall quantifies the number of true positives out of all positive predictions that could have been conducted, including true positives and false negatives, which are the ground truth IOB labels of each toponym to be identified. The F-score can be understood as a harmonic mean of Precision and Recall, combining both of them into a single score that captures both properties.

4.1 Baseline comparisons

To evaluate the performance, we modified and retrained the existing multilingual NER tools or multilingual geoparsing workflows in other languages and tested them on the Twitter dataset that we collected to compare with our proposed TRM model.

The cross-lingual NER tool proposed by Murthy et al. [74] was initially designed for NER tasks. It combines a character-level embedding processed by a convolutional neural network together with Bilbowa bilingual word embeddings [40] into a BiLSTM-CRF network to recognise toponyms and other entities. The original model was trained and tested in English, Spanish, Dutch and German. However, because the official toolkit of Bilbowa bilingual word embeddings[12] has no official implementation for the English-Finnish language pair, and further exploration on this word embedding is beyond the scope of this paper, we replaced such an embedding layer with MUSE. Therefore, in the reproduced model, it is consisted of MUSE and a character embedding layer together with a BiLSTM-CRF network. Although there are differences between how each word is embedded by a convolutional neural network as a character embedding layer and a BiLSTM layer as adopted in TRM [69], such a re-designed cross-lingual NER model can be seen as a simplified version of TRM without BERT embedding layer. We retrained such a model on our bilingual dataset (see Section 3.2.1) so that it can directly perform toponym recognition in both English and Finnish, and we compared its performance with TRM tested with the prepared Twitter dataset (see Section 3.2.2).

[9] https://colab.research.google.com/
[10] https://www.nvidia.com/en-us/
[11] https://github.com/PengyuanLiu1993/Bilingual-TRM
[12] https://github.com/gouwsmeister/bilbowa

As mentioned in Section 2, LanguageBridge [16] is a geoparsing workflow based on machine translation, translating other languages to English. The original framework was designed and tested with Chinese, Arabic, and English, yet it is possible to extend to other languages. Their framework is designed following the overall steps of *Machine translation - Word alignment - Conditional Random Field toponym recognition*. However, we soon found that their proposed word alignment step based on Fast Align [31] did not apply to our task because it requires an additional large Finnish-English dataset to train the word alignment tool. Thus, this step of word alignment is removed in the reproduced code. We used Google Translation API[13] as the machine translation tool and a trained English Conditional Random Field-based toponym recognition tool (trained with 300 WNUT 2017 data as introduced in Section 3.2.1) to recognise toponyms in our bilingual Twitter dataset.

Soon after proposing BERT [30], Google research introduced a multilingual version of BERT capable of working with more than 100 languages[14]. It can be applied directly to many down-stream tasks such as NER [9] which also identifies toponyms that are concerned with geographic entities, or it can be straightforwardly retrained on our proposed toponym recognition task. We firstly adopt the cased-version multilingual BERT developed by Burtsev et al. [13] as a NER tool to identify toponyms in the text. As a NER tool, it recognises up to 18 entities [103], including PERSON, NORP (nationalities or religious or political groups), ORG (organisations such as companies, agencies, institutions, etc.), LOC (non-GPE locations, mountain ranges, bodies of water), GPE (countries, cities, states), DATE, MONEY, FAC (buildings, airports, highways, bridges, etc.), PRODUCT, EVENT, WORK_OF_ART, LAW, LANGUAGE, TIME, PERCENT, QUANTITY, ORDINAL, and CARDINAL. One can choose to keep only LOC in the output; or, for better coverage of entities (e.g., to include locations such as institutions, schools), both LOC, FAC, GPE and ORG can be retained. As pointed by Wang et al. [101], keeping only LOC in the output may exclude other phrases that are about locations (e.g., cities, countries); however, retaining all possible entity types mentioned above will possess phrases that are not necessarily location-relevant. Taking an artificial text "Running a marathon in Helsinki, hosted by Suomen Urheiluliitto" as an example, if keeping both LOC, FAC, GPE and ORG as toponyms, "Helsinki" would be correctly identified but "Suomen (Finland/Finland's/Finnish) Urheiluliitto" as an organisation would be mistakenly understood as a toponym. Such a tricky design choice highlights the problem in using a general NER tool for toponym recognition. Following Wang et al. [101], we tested two versions of the multilingual BERT as the off-the-shelf NER tools; one version recognises phrases that are location-indicative using LOC only. In contrast, the other version follows a broader definition of location by including many entity types (i.e., LOC, FAC, GPE and ORG) that may be related to locations. Additionally, we retrained the multilingual BERT directly to our downstream toponym recognition task using the data introduced in Section 3 and compare its performance with TRM.

As shown in Table 2, our *TRM* model outperforms all baseline models introduced before in *Precision*, *Recall* and *F1-score*. As a current state-of-the-art non-deep learning-based multilingual geoparsing tool, the performance of the re-designed LanguageBridge was much worse compared to TRM. This is partially because we removed the crucial step *word alignment* of this workflow, but more importantly, the robustness of this framework is heavily dependent on the quality of machine translation. Such a point is also stated by the pre-

[13]https://cloud.google.com/translate/

[14]https://github.com/google-research/bert/blob/master/multilingual.md

vious studies on this workflow [17], that is, the higher the quality of the translation, the more precise the geoparsing process is. However, the translation quality from Finnish to English with Google translation API is not accurate enough. According to Aiken [2], comparing with Chinese and Arabic, which are used as the results' reports in Chen et al. [16], the quality of Finnish translation is much lower than the other two languages. Although the Conditional Random Field trained specifically on Twitter data can address and solve some ambiguities and irregularities in the translated text, the workflow of LanguageBridge proved to be less robust when compared to other neural network-based approaches.

Table 2: Performance of TRM and baseline models on sport and exercise-related tweets (best results reported, "∼" symbol in the table denotes for the approximations of the reported numbers; "–" symbol indicates that the corresponding numbers are not reported in the original papers or reports or published codes, and if numbers are presented in this table, they are reported from our reproduced models).

Test data	Models	Number of Parameters	Precision	Recall	F1-score
Bilingual Twitter data (1500 tweets, 750 in English and 750 in Finnish)	Re-trained cross-lingual NER [74]	–, ∼ 1.4 million	0.7467	0.6531	0.6928
	Cased multilingual BERT NER (broad location) [13]	∼ 110 million	0.8002	0.6046	0.7018
	Cased multilingual BERT NER (narrow location) [13]	∼ 110 million	0.5792	0.4541	0.5065
	Re-designed LanguageBridge [16]	–	0.6912	0.5983	0.6536
	Re-trained cased multilingual BERT	∼ 180 million	0.8019	0.7329	0.7659
	TRM	∼ 182 million	**0.8125**	**0.7335**	**0.7710**
English-only Tweets (750 tweets)	Standard NER (narrow location)	–	0.7982	0.6273	0.6831
	Standard NER (broad location)	–	0.7528	0.6547	0.7001
	NeuroTPR [101]	∼ 19 million	0.8197	**0.7943**	**0.8021**
	TRM	∼ 182 million	0.8263	0.7610	0.7835
	Re-designed TRM	∼ 181 million	**0.8371**	0.7732	0.7923
Finnish-only Tweets (750 tweets)	Re-trained FinBERT [99]	∼ 110 million	**0.8024**	**0.7542**	**0.7798**
	TRM	∼ 182 million	0.7362	0.6239	0.6831

It is interesting to see the multilingual BERT NER (*cased multilingual BERT NER, broad location*) that keeps various tags (i.e., LOC, FAC, GPE and ORG) achieves very close performance in *Precision* compared to TRM, which shows the strong capability of BERT as a NER tool to recognise toponyms. However, low performance in *Recall* indicates many correct locations are not identified. This is because the NER model [13] managed to identify city names or countries (i.e., Helsinki or Finland that commonly exists in tweets), while it failed to identify more fine-grained toponyms such as street names, especially when the names of the streets are in Finnish. Meanwhile, the multilingual BERT NER that keeps LOC only (*cased multilingual BERT NER, narrow location*) achieves the worst performance among all models since many toponyms are classified as other entities; thus, merely relying on LOC tags in the NER model is not enough for toponym recognition tasks.

We also introduced two neural network-based approaches using cross-lingual NER [74] and multilingual BERT (*re-trained cased multilingual BERT* in Table 2), which were retrained on our prepared training data and tested with the defined toponym recognition task. The BERT was trained with a BiLSTM as the backend and combined with a CRF layer to produce tags for the toponyms. As shown in Table 2, the re-trained multilingual BERT outperformed the retrained cross-lingual NER, which demonstrates and agrees with most existing

studies that BERT can achieve state-of-the-art performance that superior to many previously developed models [93]. Despite slight lower *Precision*, *Recall* and *F1-score*, re-trained cased multilingual BERT achieved a performance level that was close to our TRM. A closer examination of this comparison results shows that the combination of the character-level embedding layer demonstrates a stronger ability to capture misspelling errors in the text.

4.2 Ablation studies

Previous experiments showed the strong capabilities of TRM in addressing bilingual toponym recognition tasks. As a bilingual model, it provides the ability to work in a single language usage context. To demonstrate the flexibility of TRM, we designed individual experiments to compare our proposed model with a set of state-of-the-art models in English-only and Finnish-only settings with the prepared Twitter dataset.

4.2.1 English-only tweets

There is a large body of literature on toponym recognition tools developed for English because it is a high-resourced language. We compared TRM with the off-the-shelf Stanford NER tool (two versions), and a deep learning-based model (*NeuroTPR*) designed explicitly for toponym recognition in English social media text. We tested the models using 750 English tweets introduced in Section 3.2.2, and the results are summarised in Table 2. We adopted the widely used three-class Stanford NER tool in the experiments that produced PERSON, ORGANISATION, and LOCATION as identified entities. *Narrow location* for Stanford NER tool means the model only keeps LOCATION as the output; meanwhile, *broad location* is that Stanford NER tool keeps all entity types that might be related to locations (i.e., LOCATION and ORGANISATION).

As shown in Table 2, despite both TRM and NeuroTPR outperforming the Stanford NER tool, such a classic NER tool still demonstrates its effectiveness recognising toponyms. NeuroTPR achieves the highest performance in *Recall* and *F1-score*, which indicates the model is more robust in recognising correct toponyms comparing with our proposed TRM. It is interesting from a linguistic point of view that a bilingual or multilingual model might not be able to understand linguistic features and words better than language-specific models. To further address such a point, we proposed an additional experiment tested with English-specific TRM (*re-designed TRM* in Table 2). We replaced the static multilingual word embedding layer as introduced in Figure 1 with pre-trained English Twitter-specific word embeddings [38] to represent the words in a tweet. We also replaced multilingual BERT layer with a cased English BERT ("bert-large-cased" from Wolf et al. [106]; see also the online document on the *Hugging Face*[15] website) to extract embeddings from the text. The performance of such an English-specific model is superior to the original TRM on toponym recognition in English. However, although the performance of both model is similar, NeuroTPR seems still slightly better according to *Recall* and *F1-score*. As mentioned in Section 3.1, our model was inspired by the success of NeuroTPR [101] and it has a similar but distinctive model architecture. One of the main differences is that we dropped the part-of-speech tagging embedding layer entirely. The part-of-speech tagging embedding layer informs the model about the word type (noun, verb, adjective, preposition, etc.) so that NeuroTPR can learn how location-indicative words are used and placed in a sentence (e.g.,

[15]https://huggingface.co/transformers/pretrained_models.html

a preposition phrase is often followed by a location). However, in the context of bilingual tasks, Finnish and English have different grammars; thus, the different use patterns of the words might confuse the model and decrease the performance. Therefore, we removed that layer to reduce the impact of English and Finnish's distinctive linguistic features during the training. Nevertheless, the results in Table 2 suggest that such a part-of-speech tagging embedding layer can benefit the model performance when applied to language-specific tasks.

4.2.2 Finnish-only tweets

We retrained the FinBERT (Finnish BERT model) [99] with our prepared training data (Finnish-only) introduced in Section 3.2 as a comparison to our proposed TRM. Like the results of the experiments presented in Section 4.2.1, language-specific model FinBERT outperformed our TRM. In terms of comparisons between language-specific models and our proposed bilingual model, the performance difference between FinBERT and TRM was much more significant than the difference between TRM and NeuroTPR. This is because of the more abundant English training data (3300 sentences) compared with the Finnish training data (1587 sentences), as we introduced in Section 3. Thus, TRM seemed to learn limited linguistic features in Finnish compared to English.

The ablation studies reinforced the concept that "multilingual is not enough" [99]. That is, in language-specific use context, designing and training a deep learning model that can specifically focus on the language's linguistics and word patterns would perform better than applying a bi- or multilingual framework.

5 Showcase study for sport and exercise-related tweets

In the previous sections, we have introduced our proposed TRM for toponym recognition tasks from social media messages. This section presents a showcase study using "real-world" social media data (i.e., tweets) collected by TRM. There are two primary research objectives of this showcase study. Our first aim is to validate the usability of the TRM, and show how the model can be generalised to practical research tasks and benefit the data collection process. Our second aim is to explore the potential of social media as a source of information on sports and exercise in cities based on aggregated spatial data. This showcase study operationalised the concept mentioned in Section 2 that place representation described by UGC is the amount and type of information available in an area [7,8] to quantify the spatial relationships between tweets and socio-economic variables. Such a showcase study contributes to our understanding of the informational dimension of place by comparing the spatial Twitter data with a specific focus on sport and exercise-related activities to their socio-economic and physical infrastructural (i.e., sport facilities) context that impacts the production process.

5.1 Usability of TRM

Figure 3 is a showcase of using our proposed TRM as a toponym recognition tool together with a Google Geocoding API[16] as a toponym resolution tool to perform geoparsing. Note that one can replace Google Geocoding API with other tools, such as geocoding function

[16]https://developers.google.com/maps/documentation/geocoding/overview

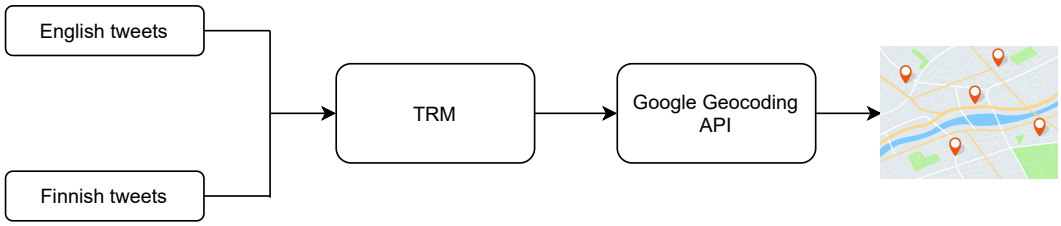


Figure 3: A showcase of the potential geoparsing workflow using TRM.

from GeoPandas[17], or other services. It is worth noting that these services do not automatically perform place name disambiguation since they don't know the contexts under which these toponyms are mentioned. However, as the showcase study in this section will focus on the city of Helsinki, the names of streets or places mentioned in the text would not be highly ambiguous; thus, it will be acceptable to use Google Geocoding API for the preliminary results showcase.

We followed the same matching-on-form approach as described in Section 3.2.2 to collect tweets that mention sport and exercise-related activities in English and Finnish from the original Twitter dataset. Note that a similar issue remains in the dataset; there are still many tweets that are "noisy" because the matching-on-form approach is relatively naive in filtering out sport and exercise-related tweets. However, such an issue is beyond the scope of this paper, and one to be investigated and addressed in our future research. It is also worth noting that we decided to exclude hashtags in the analysis as a designing choice. Hashtag is often considered a very useful self-reporting tool for users to locate themselves [48, 82], however, through our manual inspection of the tweets, the locations reported using hashtags are often too general, and the most common seen hashtags used in tweets refers to the city name (i.e., "Helsinki") or the country name (i.e., "Finland"). This is may due to the fact that the more general hashtags users are using for location self-reporting the more attention are likely to be gotten on the Internet. As one of the objectives of this study is to collect as fine-grained toponyms in the text as possible, we exclude hashtags of tweets in this analysis.

The next step was to use TRM to recognise the candidate location-indicative words (i.e., toponyms) in the tweets. In our experiments, we compare the time cost using TRM that is directly applied to both English and Finnish tweets with the time cost applying language-specific models to the two languages separately (FinBERT for Finnish tweets and NeruoTPR for English tweets) in chronological order. The TRM demonstrates a 32.2% time efficiency improvement, which indicates a bilingual model can simplify the data collection process. Note that we explicitly excluded tweets that only mention toponyms as the city name of Helsinki or the country name of Finland. Therefore, the tweets left in the dataset are the tweets with toponyms on the resolution at neighbourhood scales (e.g., names of postcode areas) or more fine-grained point-of-interest scales (e.g., building, street names and parks). Such a step is to demonstrate the usability of TRM in recognising toponyms at fine-grained resolutions. As a result of this step, we got 56,784 tweets and their corresponding toponyms. Meanwhile, our TRM also demonstrates a strong ability to handle "code-switching" in the text where both English and Finnish are presented. Taking the

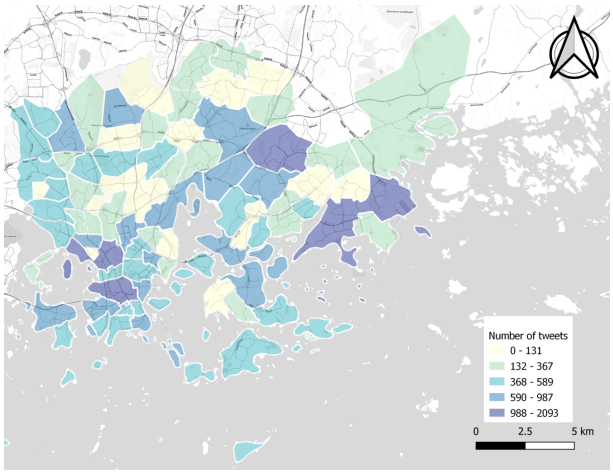
[17]https://geopandas.org/docs/user_guide/geocoding.html

sentence "Upea sää Kumpulassa (great weather in Kumpula, Kumpla is a verdant neighbourhood in Helsinki), nice walk in Helsinki University" as an example, TRM can recognise both "Kumpulassa" and "Helsinki University" as toponyms to geo-locate but NeuroTPR can only recognise "Helsinki University" and FinBERT only recognises "Kumpulassa".

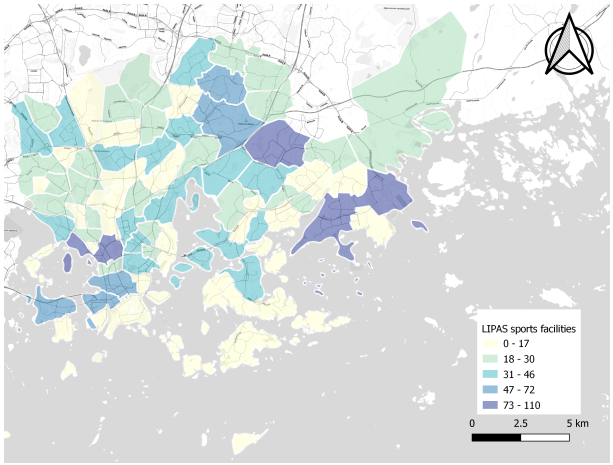
After obtaining toponyms from tweets, Google Geocoding API performed geocoding on the toponyms and converted them into geo-coordinates that can be visualised on the map. We aggregated the data into postcode areas in Helsinki. Despite the potential issues of the modifiable areal unit problem (MAUP) and uncertainties that irregular boundaries may cause [78], postcode areas are widely adopted within urban studies to visualise geographical patterns and compare places based on the data aggregated at the neighbourhood scale, and they are the ideal spatial units for this showcase study. Note that for tweets that have more than one toponyms, for example, "Rowing from Töölönlahti (Töölö bay) to Kaisaniemenlahti (Kaisaniemi bay)", we map such activity in both postcodes of where "Töölönlahti" and "Kaisaniemenlahti" are located. Figure 4 (a) presents a map visualised with the tweets collected from all of Helsinki over 84 postcode areas.

We further compared Figure 4 (a) with the data collected from LIPAS sport facility GIS-database [97]. LIPAS is Finland's national database of sports facilities and contains all sports facilities in Finland. The sport facilities and built environment can be considered as a potential physical infrastructural context that encourages sport-related content production from Twitter users when they are engaging into sport and exercise-related activities. There are three data types: point data for sports facilities like swimming halls, line data for walking, running, skiing and biking routes, and polygon data for recreational areas like national parks or natural reserves. In this showcase study, we aggregated sports facilities (point data) and the centroids of polygon data into postcode areas and presented the results in Figure 4 (b). Because parks and natural reserves widely attract various kinds of sport activities (e.g., walking, hiking, etc), such a map broadly covers both formal and informal sport activities in Helsinki. As mentioned above, LIPAS is a national database designed for registered sport facilities suitable for analysing sport activities at known locations. However, because one of the aims of this paper is to investigate all sport at both known and arbitrary locations at the same time, adopting the data of parks might be one of the few possible ways to include some facilities that can support "informal sport" in the LIPAS. Future research will require data from other sources (e.g., surveys) to include more locations that are facilitated for "informal sport". Note that our approach, which aggregated polygon data using centroids, is naive in displaying how parks and natural reserves are located in Helsinki. However, we consider more sophisticated aggregation methods (e.g., count polygon data in the corresponding postcode areas they spread across) beyond the scope of this showcase study and one to be pursued in our future research.

A visual comparison between 4 (a) and (b) indicates some correlation between the distribution of sport and exercise-related tweets and sport facilities across the city of Helsinki. Some postcode areas that have high numbers of tweets also have high numbers of sport facilities. To further model the geographies of tweets, we calculated the correlation coefficients among socio-economic variables, sport facilities and tweets at the postcode level in the next subsection.



(a) A showcase of TRM results in Helsinki.



(b) LIPAS dataset in Helsinki.

Figure 4: A showcase of TRM results and the comparison to the LIPAS dataset. Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

5.2 The geographies of sport and exercise-related tweets

Kahma (2012) [55] identified age, education background, and level of income conveyed some crucial differences in sporting activity participation in Finland, based on survey questionnaires. For example, taking income into account revealed that those with higher income were more likely to participate in sports. To investigate whether the digitised place might reveal different aspects of sporting participation, we also included population in different age groups (18-54 yrs.), education and income categories in our paper as socio-economic variables collected from Finnish Statistics [76] to compare the distribution with Twitter data. In the remainder of this section, we use the data mentioned above to quan-

tify the spatial relationships between sport and exercise-related tweets and socio-economic variables, based on the ties-adjusted rank correlation coefficient (Spearman's ρ).

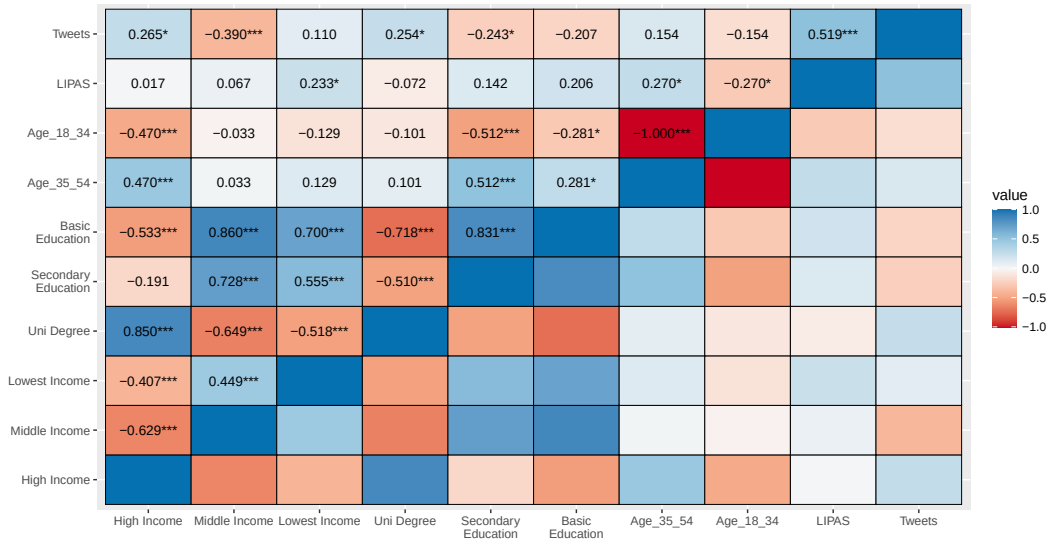


Figure 5: Spearman's correlation coefficients between demographic and UGC variables (84 postcode areas). Data normalised per 100 people. Significance levels: $*p<0.05$; $**p<0.01$; $***p<0.001$.*LIPAS* in the figure refers to the distribution of sport facilities recorded in the LIPAS system.

Figure 5 demonstrates correlation coefficients between all relevant variables introduced above at the postcode area level. When normalising data per 100 people, sport and exercise-related tweets demonstrate a significant positive correlation with the LIPAS sport facilities (0.519, $p<0.001$). Such a strong correlation indicates that the built physical environment (i.e., sport facilities) and parks for recreational activities contribute to most sport and exercise activities on Twitter. Although the correlation analysis suggests social media users are more likely to post their sport and exercise content when they are using formal physical sports facilities or exercising in parks, it is worth noting that the spatial units (i.e., postcode areas) we adopted in this study are too large to observe detailed correlation between tweets and sporting facilities. Spatial units at finer resolutions (see discussion in Section 6) are needed for our future research.

Meanwhile, the distribution of tweets demonstrates significant correlations with population groups that have high education background (university degrees) (0.265, $p<0.05$) and high income (0.254, $p<0.05$). Such a finding echos many existing studies that Twitter tends to be more representative of wealthier urban areas, inhabited by more educated populations than average [7, 8, 87]. Interestingly, the sports facilities in the LIPAS system seem to have no significant correlations with most socio-economic variables, suggesting an overall success of the government's agenda in promoting equal accessibility to sport facilities [94]. However, the accessibility of sport facilities can be improved for young adults because they demonstrate a significant negative correlation with the population at the age group between 18 and 34 (-0.270, $p<0.05$).

The correlation comparison presented above indicates that despite the overall equal accessibility of sport facilities in Helsinki for different population groups, Twitter conveys that sport and exercise-related tweets are more representative for the areas with population that has high income and good education background, indicating population in those areas is more willing in participating in sporting activities. Such an outcome supports the research finding proposed by Kahma (2012) [55] using more abundant user-generated content (i.e., tweets) online, suggesting potential segregation in the sport participation due to different socio-economic backgrounds of the residents in the city. It is important to note that following similar research methods in Ballatore and De Sabbata (2020) [8], the agnostic relationship between the geographies of content (i.e., where tweet is) and residential geographies of the content producers (i.e., where users live) is out of the scope for this showcase study but one limitation to overcome in our future research.

6 Conclusion and outlook

The primary contribution of this paper is TRM, a novel approach to the exploratory analysis of sport and exercise-related social media content. TRM is capable of identifying location-indicative words (i.e., toponyms) straightforwardly from the bilingual text so that it can simplify the data collection process and benefit studies of urban places using the collected spatial data. An important advantage of TRM is its ability to recognise many fine-grained toponyms, which off-the-shelf NER tools are often struggled with in bilingual context usage. Meanwhile, TRM demonstrates a strong ability to address the issue of "code-switching" that is commonly seen in social media texts, leading to broader coverage of spatial data collection when two languages are presented in one piece of text. This paper also contributes a quantitative framework to investigate "informal activities" that could potentially be restricted to specific locations and facilities if users choose to post online. Data and locations retrieved from social media using our TRM can be served as supplementary data for the national database of sport facilities (i.e., LIPAS) to include more possible locations that are used for sporting activities. It is important to emphasise that despite that our model being primarily designed for the toponym recognition task targeting on tweets posted in Finland, TRM can easily be adapted and extended to other languages or further developed as a multilingual tool with carefully prepared training data. We additionally presented a showcase study demonstrating the usefulness of our TRM using tweets collected at the level of the city of Helsinki, and it can be integrated easily into a geoparsing workflow.

We hope to pursue this research in several directions in our future studies. First, as mentioned in Section 3.2.2, we introduced a matching-on-form approach to collect sport and exercise-related tweets. Although the matching-on-form approach is easy to implement, such an approach eventually results in a large amount of noise in the dataset, which requires extra human interventions to clean the data. Our future research can integrate TRM with a text classification tool (e.g., [32, 56, 92]) as a workflow to filter out sport and exercise-related messages with toponyms to be identified more precisely. Going one step further, we could use toponyms recognised by TRM as geographic information to construct a graph-based representation of where the sport and exercise-related activities happen. Then we can utilise such a representation to classify tweets further into fine-grained categories (e.g., swimming, jogging, etc.) [68] or study place characteristics [109]. Second,

in Section 5, the geoparser adopted Google Geocoding API as a toponym resolution tool. However, street names are highly ambiguous. Google Geocoding API does not perform automatic place name disambiguation; such an issue increases the challenge when applying the geoparser to larger-scale geographic regions such as the national scale Finland. Taking "Välskärinkatu", a randomly picked street name as an example, if users do not specify the city where Välskärinkatu is in their post, all the tweets mention this street will be automatically geocoded to a coordinates pair in Helsinki. However, Välskärinkatu is also a street name in many other cities, leading to further uncertainties. Future research may require us to combine place name disambiguation techniques [54] to reduce the impact of such an issue. Third, many existing geoparsers geocode a toponym with a single pair of coordinates in the form of point; however, research sometimes requires toponyms to be geolocated in other forms of spatial footprints, such as lines and polygons. For example, in the study presented in Section 5, for a sentence such as "Jogging in Fredrikinkatu", the line representation might be a better choice since Fredrikinkatu as a street crosses two postcode areas. Merely counting based on the point data located in one area in such a case will raise uncertainties on the research output.

This paper also provides a showcase study that compares the distribution of sport and exercise-related tweets and physical sports facilities in the LIPAS dataset with socio-economic variables. The aim of the study was to understand how social media and user-generated content can be used as a source of information on sport and exercise in cities. We outlined two interesting findings based on the case study presented in Section 5.2: 1). the built physical environment (i.e., sport facilities) and parks for recreational activities contributes to most sport and exercise activities on Twitter; 2). sport and exercise-related tweets tend to be more representative of wealthier urban areas, inhabited by more educated populations than average.

However, we would like to highlight several challenges in showcase study that will be pursued in our future research. First, the conclusions are drawn from a relatively "noisy" Twitter dataset, and such a dataset still has a large amount of "noises" that are not cleaned after the matching-on-form approach introduced in Section 3.2. Our future research will perform analysis on a cleaner Twitter dataset to draw more concrete conclusions. Second, in the case study, we illustrate that parks and built physical environments are likely to have sport and exercise-related activities. Further investigations into land type comparisons might provide a better indication of what types of land use correlate to a higher number of sporting activities and sports facilities. Third, the scale of postcode areas we adopted is too large to observe detailed correlations between Twitter and the distribution of sports facilities. In our future research, we will investigate such correlations at finer spatial resolutions (e.g., 250 metre $\times$ 250 metre spatial grids). Fourth, this study focused on a single city. A comparative approach with other cities and rural or semi-rural areas or not-so-dense near urban areas, in Finland and elsewhere, is necessary to observe the geographical variation in the relationships between the place and the sport and exercise-related activities that are carried out. Moreover, our analysis needs to include more land use and census data to capture the role of urban function and mass population and other socio-economic factors (e.g., occupations, as in [55]), prominent in many data-rich areas in central Helsinki. Fifth, investigations of the impact of temporal-dimension of social media activities are currently missing. Future research will investigate how places are represented and perceived by sport and exercise-related content generated at different times of a day or a year by performing temporal analysis.

In summary, this paper provides new tools in geographical studies to collect spatial data. It explores how sport contributes to understanding space and place by bridging user-generated content from social media platforms and the development of a quantitative artificial intelligence method. We consider our proposed TRM as a valuable addition to the discipline of sports geography, and can be implemented more broadly in geographical information retrieval tasks or other geographical studies that require collecting user-generated content from online platforms. In the development of quantitative models and algorithms which incorporate user-generated content as an essential source of information to study the relationships between users' activities (including sport) and place, we are ultimately able to identify ways that can benefit our understanding of the online socio-spatial process in the urban context and its impact on the physical environment we are living in.

Acknowledgments

This study is a part of the "Equality in suburban physical activity environments, YLLI" research project (in Finnish: Yhdenvertainen liikunnallinen lähiö, YLLI). The project is being financed by the research program about suburban in Finland "Lähiöohjelma 2020-2022" coordinated by the Ministry of Environment (grant recipient: Dr. Petteri Muukkonen). In addition, we would like to thank the Digital Geography Lab, University of Helsinki, for its data and database support for this study. In addition, we are grateful to researchers and colleagues in the Digital Geography lab for their encouragements and valuable comments.

References

[1] ABERNATHY, D. *Using geodata and geolocation in the social sciences: Mapping our connected world*. Sage, 2016.

[2] AIKEN, M. An updated evaluation of google translate accuracy. *Studies in linguistics and literature 3*, 3 (2019), 253–260.

[3] AKBIK, A., BERGMANN, T., BLYTHE, D., RASUL, K., SCHWETER, S., AND VOLLGRAF, R. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)* (2019), pp. 54–59.

[4] ALEX, B. Geoparsing English-language text with the Edinburgh geoparser. *Programming Historian* (2017).

[5] BACH, L. Sports without facilities: the use of urban spaces by informal sports. *International review for the sociology of sport 28*, 2-3 (1993), 281–296.

[6] BALE, J., AND DEJONGHE, T. Sports geography: an overview. *Belgeo. Revue belge de géographie*, 2 (2008), 157–166.

[7] BALLATORE, A., AND DE SABBATA, S. Charting the geographies of crowdsourced information in Greater London. In *The Annual International Conference on Geographic Information Science* (2018), Springer, pp. 149–168.

[8] BALLATORE, A., AND DE SABBATA, S. Los Angeles as a digital place: The geographies of user-generated content. *Transactions in GIS 24*, 4 (2020), 880–902.

[9] BAUMANN, A. Multilingual language models for named entity recognition in German and English. In *Proceedings of the Student Research Workshop Associated with RANLP 2019* (Varna, Bulgaria, Sept. 2019), INCOMA Ltd., pp. 21–27. 10.26615/issn.2603-2821.2019_004.

[10] BILLAUDEAU, N., OPPERT, J.-M., SIMON, C., CHARREIRE, H., CASEY, R., SALZE, P., BADARIOTTI, D., BANOS, A., WEBER, C., AND CHAIX, B. Investigating disparities in spatial accessibility to and characteristics of sport facilities: Direction, strength, and spatial scale of associations with area income. *Health & place 17*, 1 (2011), 114–121.

[11] BOJKOVSKỲ, M., AND PIKULIAK, M. STUFIIT at SemEval-2019 task 5: Multilingual hate speech detection on Twitter with MUSE and ELMo embeddings. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (2019), pp. 464–468.

[12] BRAUMÜLLER, B. Hockey im club oder skaten im park?: Eine sekundäranalyse der medikus-studie zur sozialisation in vereinsorganisierte und informelle sportsettings in abhängigkeit von sozialen, personalen und medialen ressourcen in der adoleszenz. *Sport und Gesellschaft 13*, 3 (2016), 215–249.

[13] BURTSEV, M., SELIVERSTOV, A., AIRAPETYAN, R., ARKHIPOV, M., BAYMURZINA, D., BUSHKOV, N., GUREENKOVA, O., KHAKHULIN, T., KURATOV, Y., KUZNETSOV, D., ET AL. Deeppavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations* (2018), pp. 122–127.

[14] CARTER, S., WEERKAMP, W., AND TSAGKIAS, M. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation 47*, 1 (2013), 195–215.

[15] CHACÓN-BORREGO, F., CORRAL-PERNÍA, J. A., MARTÍNEZ-MARTÍNEZ, A., AND CASTAÑEDA-VÁZQUEZ, C. Usage behaviour of public spaces associated with sport and recreational activities. *Sustainability 10*, 7 (2018), 2377.

[16] CHEN, X., GELERNTER, J., ZHANG, H., AND LIU, J. Multi-lingual geoparsing based on machine translation. *Future Generation Computer Systems 96* (2019), 667–677. https://doi.org/10.1016/j.future.2017.07.057.

[17] CHEN, X., ZHANG, H., AND GELERNTER, J. Multi-lingual geoparsing based on machine translation. *arXiv preprint arXiv:1511.01974* (2015).

[18] CHIRAZI, M. Comparative evolution of the phenomenon of geography of sports on national and global levels. *Geosport for Society 10*, 1 (2019), 7–14.

[19] CITY OF ESPOO. *Exercise Classes*, 2021. Available at https://www.espoo.fi/en-US/Culture_and_sport/Sports/Exercise_classes (Accessed July 28th, 2021).

[20] CITY OF HELSINKI. *Physical Activity Programme*, 2018. Available at https://helsinkiliikkuu.fi/en/liikkumisohjelma/ (Accessed July 28th, 2021).

[21] CLAUSNITZER, J. *Social media usage in Finland - statistics & facts*, 2021. Available at https://www.statista.com/topics/4173/social-media-usage-in-finland/ (Accessed September 27th, 2021).

[22] CLAVIO, G. *Social media and sports*. Human Kinetics Publishers, 2020.

[23] CONNEAU, A., LAMPLE, G., RANZATO, M., DENOYER, L., AND JÉGOU, H. Word translation without parallel data. *arXiv preprint arXiv:1710.04087* (2017).

[24] COOMBES, E., JONES, A. P., AND HILLSDON, M. The relationship of physical activity and overweight to objectively measured green space accessibility and use. *Social science & medicine 70*, 6 (2010), 816–822.

[25] CORNEGRUTA, S., BAKEWELL, R., WITHEY, S., AND MONTANA, G. Modelling radiological language with bidirectional long short-term memory networks. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis* (Auxtin, TX, Nov. 2016), Association for Computational Linguistics, pp. 17–27. 10.18653/v1/W16-6103.

[26] CROOKS, A., PFOSER, D., JENKINS, A., CROITORU, A., STEFANIDIS, A., SMITH, D., KARAGIORGOU, S., EFENTAKIS, A., AND LAMPRIANIDIS, G. Crowdsourcing urban form and function. *International Journal of Geographical Information Science 29*, 5 (2015), 720–741.

[27] DEL GRATTA, R., GOGGI, S., PARDELLI, G., AND CALZOLARI, N. The LRE map: what does it tell us about the last decade of our field? *Language Resources and Evaluation 55*, 1 (2021), 259–283.

[28] DELOZIER, G., BALDRIDGE, J., AND LONDON, L. Gazetteer-independent toponym resolution using geographic word profiles. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2015), vol. 29.

[29] DERCZYNSKI, L., XU, W., RITTER, A., AND BALDWIN, T. Proceedings of the 3rd workshop on noisy user-generated text. In *Proceedings of the 3rd Workshop on Noisy User-generated Text* (2017).

[30] DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[31] DYER, C., CHAHUNEAU, V., AND SMITH, N. A. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2013), pp. 644–648.

[32] EDWARDS, A., CAMACHO-COLLADOS, J., DE RIBAUPIERRE, H., AND PREECE, A. Go simple and pre-train on domain-specific corpora: On the role of training data for text classification. In *Proceedings of the 28th International Conference on Computational Linguistics* (Barcelona, Spain (Online), Dec. 2020), International Committee on Computational Linguistics, pp. 5522–5529. 10.18653/v1/2020.coling-main.481.

[33] ELETA, I., AND GOLBECK, J. Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior 41* (2014), 424–432.

[34] FREIRE, N., BORBINHA, J., CALADO, P., AND MARTINS, B. A metadata geoparsing system for place name recognition and resolution in metadata records. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries* (2011), pp. 339–348.

[35] FRIDBERG, T. Sport and exercise in Denmark, Scandinavia and Europe. *Sport in Society 13*, 4 (2010), 583–592.

[36] GELERNTER, J., AND BALAJI, S. An algorithm for local geoparsing of microtext. *GeoInformatica 17*, 4 (2013), 635–667.

[37] GELERNTER, J., AND MUSHEGIAN, N. Geo-parsing messages from microtext. *Transactions in GIS 15*, 6 (2011), 753–773.

[38] GODIN, F., VANDERSMISSEN, B., DE NEVE, W., AND VAN DE WALLE, R. Multimedia lab@ acl wnut ner shared task: Named entity recognition for Twitter microposts using distributed word representations. In *Proceedings of the workshop on noisy user-generated text* (2015), pp. 146–153.

[39] GOODCHILD, M. F. Citizens as sensors: the world of volunteered geography. *GeoJournal 69*, 4 (2007), 211–221.

[40] GOUWS, S., BENGIO, Y., AND CORRADO, G. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning* (Lille, France, 07–09 Jul 2015), F. Bach and D. Blei, Eds., vol. 37 of *Proceedings of Machine Learning Research*, PMLR, pp. 748–756.

[41] GRITTA, M., PILEHVAR, M. T., AND COLLIER, N. Which Melbourne? augmenting geocoding with maps. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2018), pp. 1285–1296.

[42] GRITTA, M., PILEHVAR, M. T., LIMSOPATHAM, N., AND COLLIER, N. What's missing in geographical parsing? *Language Resources and Evaluation 52*, 2 (2018), 603–623.

[43] HEIKINHEIMO, V., MININ, E. D., TENKANEN, H., HAUSMANN, A., ERKKONEN, J., AND TOIVONEN, T. User-generated geographic information for visitor monitoring in a national park: A comparison of social media data and visitor survey. *ISPRS International Journal of Geo-Information 6*, 3 (2017), 85.

[44] HEIKINHEIMO, V., TENKANEN, H., BERGROTH, C., JÄRV, O., HIIPPALA, T., AND TOIVONEN, T. Understanding the use of urban green spaces from user-generated geographic information. *Landscape and Urban Planning 201* (2020), 103845.

[45] HERBERT, D., AND THOMAS, C. *Cities in space: city as place*. Routledge, 2013.

[46] HERDAĞDELEN, A. Twitter n-gram corpus with demographic metadata. *Language resources and evaluation 47*, 4 (2013), 1127–1147.

[47] HIIPPALA, T., VÄISÄNEN, T. L. A., TOIVONEN, T., JÄRV, O., ET AL. Mapping the languages of Twitter in Finland: Richness and diversity in space and time. *Neuphilologische Mitteilungen* (2020).

[48] HOANG, T. B. N., AND MOTHE, J. Location extraction from tweets. *Information Processing & Management 54*, 2 (2018), 129–144.

[49] HU, Y., MAO, H., AND MCKENZIE, G. A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements. *International Journal of Geographical Information Science 33*, 4 (2019), 714–738.

[50] HU, Y., AND WANG, R.-Q. Understanding the removal of precise geotagging in tweets. *Nature Human Behaviour* (2020), 1–3.

[51] HUANG, Q., AND WONG, D. W. Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science 30*, 9 (2016), 1873–1898.

[52] ILIEVA, R. T., AND MCPHEARSON, T. Social-media data for urban sustainability. *Nature Sustainability 1*, 10 (2018), 553–565.

[53] JOSHI, V., PETERS, M., AND HOPKINS, M. Extending a parser to distant domains using a few dozen partially annotated examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2018), pp. 1190–1199.

[54] JU, Y., ADAMS, B., JANOWICZ, K., HU, Y., YAN, B., AND MCKENZIE, G. Things and strings: improving place name disambiguation from short texts by combining entity co-occurrence with topic modeling. In *European Knowledge Acquisition Workshop* (2016), Springer, pp. 353–367.

[55] KAHMA, N. Sport and social class: The case of Finland. *International Review for the Sociology of sport 47*, 1 (2012), 113–130.

[56] KANT, N., PURI, R., YAKOVENKO, N., AND CATANZARO, B. Practical text classification with large pre-trained language models. *arXiv preprint arXiv:1812.01207* (2018).

[57] KARIMZADEH, M. Performance evaluation measures for toponym resolution. In *Proceedings of the 10th workshop on geographic information retrieval* (2016), pp. 1–2.

[58] KARIMZADEH, M., HUANG, W., BANERJEE, S., WALLGRÜN, J. O., HARDISTY, F., PEZANOWSKI, S., MITRA, P., AND MACEACHREN, A. M. GeoTxt: a web API to leverage place references in text. In *Proceedings of the 7th workshop on geographic information retrieval* (2013), pp. 72–73.

[59] KARIMZADEH, M., PEZANOWSKI, S., MACEACHREN, A. M., AND WALLGRÜN, J. O. GeoTxt: A scalable geoparsing system for unstructured text geolocation. *Transactions in GIS 23*, 1 (2019), 118–136.

[60] KARUSISI, N., THOMAS, F., MÉLINE, J., AND CHAIX, B. Spatial accessibility to specific sport facilities and corresponding sport practice: the record study. *International Journal of Behavioral Nutrition and Physical Activity 10*, 1 (2013), 1–10.

[61] KINGMA, D. P., AND BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[62] KITTUR, A., CHI, E. H., AND SUH, B. What's in wikipedia? mapping topics and conflict using socially annotated category structure. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2009), pp. 1509–1512.

[63] KOCH, N. Sports and the city. *Geography Compass 12*, 3 (2018), e12360.

[64] KOSKI, P., AND LÄMSÄ, J. Finland as a small sports nation: socio-historical perspectives on the development of national sport policy. *International Journal of Sport Policy and Politics 7*, 3 (2015), 421–441.

[65] LAMPLE, G., BALLESTEROS, M., SUBRAMANIAN, S., KAWAKAMI, K., AND DYER, C. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (San Diego, California, June 2016), Association for Computational Linguistics, pp. 260–270. 10.18653/v1/N16-1030.

[66] LANSLEY, G., AND LONGLEY, P. A. The geography of Twitter topics in London. *Computers, Environment and Urban Systems 58* (2016), 85–96.

[67] LIEBERMAN, M. D., SAMET, H., AND SANKARANARAYANAN, J. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *2010 IEEE 26th international conference on data engineering (ICDE 2010)* (2010), IEEE, pp. 201–212.

[68] LIU, P., AND DE SABBATA, S. A graph-based semi-supervised approach to classification learning in digital geographies. *Computers, Environment and Urban Systems 86* (2021), 101583. https://doi.org/10.1016/j.compenvurbsys.2020.101583.

[69] LUAN, Y., AND LIN, S. Research on text classification based on CNN and LSTM. In *2019 IEEE international conference on artificial intelligence and computer applications (ICAICA)* (2019), IEEE, pp. 352–355.

[70] LUOMA, J., OINONEN, M., PYYKÖNEN, M., LAIPPALA, V., AND PYYSALO, S. A broad-coverage corpus for Finnish named entity recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference* (2020), pp. 4615–4624.

[71] MÄKINEN, K., AND TYRVÄINEN, L. Teenage experiences of public green spaces in suburban Helsinki. *Urban forestry & urban greening 7*, 4 (2008), 277–289.

[72] MANNING, C. D., SURDEANU, M., BAUER, J., FINKEL, J. R., BETHARD, S., AND MCCLOSKY, D. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (2014), pp. 55–60.

[73] MILLER, H. J., AND GOODCHILD, M. F. Data-driven geography. *GeoJournal 80*, 4 (2015), 449–461.

[74] MURTHY, R., KHAPRA, M., AND BHATTACHARYYA, P. Sharing network parameters for crosslingual named entity recognition. *arXiv preprint arXiv:1607.00198* (2016).

[75] NICOLSON, M. Sport-for-inclusion in Helsinki: A field of tension between policy-makers and practitioners.

[76] PAAVO POSTAL CODE AREA STATISTICS. *Paavo postal code area statistics 2021*, 2021. Available at https://www.stat.fi/tup/paavo/paavon_aineistokuvaukset_en.html (Accessed May 28th, 2021).

[77] PASZKE, A., GROSS, S., MASSA, F., LERER, A., BRADBURY, J., CHANAN, G., KILLEEN, T., LIN, Z., GIMELSHEIN, N., AND ANTIGA, L. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems 32* (2019), 8026–8037.

[78] PFEFFER, K., DEURLOO, M. C., AND VELDHUIZEN, E. M. Visualising postcode data for urban analysis and planning: the amsterdam city monitor. *Area 44*, 3 (2012), 326–335.

[79] PORTAZ, M., RANDRIANARIVO, H., NIVAGGIOLI, A., MAUDET, E., SERVAN, C., AND PEYRONNET, S. *Image search using multilingual texts: a cross-modal learning approach between image and text*. PhD thesis, qwant research, 2019.

[80] PURVES, R. S., CLOUGH, P., JONES, C. B., HALL, M. H., AND MURDOCK, V. Geographic information retrieval: Progress and challenges in spatial search of text. *Foundations and Trends in Information Retrieval 12*, 2-3 (2018), 164–318.

[81] PURVES, R. S., WINTER, S., AND KUHN, W. Places in information science. *Journal of the Association for Information Science and Technology 70*, 11 (2019), 1173–1182.

[82] REELFS, H., MOHAUPT, T., HOHLFELD, O., AND HENCKELL, N. Hashtag usage in a geographically-local microblogging app. In *Companion Proceedings of The 2019 World Wide Web Conference* (2019), pp. 919–927.

[83] ROBERTS, H., SADLER, J., AND CHAPMAN, L. Using Twitter to investigate seasonal variation in physical activity in urban green space. *Geo: Geography and Environment 4*, 2 (2017), e00041.

[84] ROCHE, S. Geographic information science II: Less space, more places in smart cities. *Progress in Human Geography 40*, 4 (2016), 565–573.

[85] SALONEN, R.-M. Socio-economics of figure skating: Finnish figure skating families' socio-economic standing and perceptions of their child's participation in figure skating.

[86] SALZE, P., BANOS, A., OPPERT, J.-M., CHARREIRE, H., CASEY, R., SIMON, C., CHAIX, B., BADARIOTTI, D., AND WEBER, C. Estimating spatial accessibility to facilities on the regional scale: an extended commuting-based interaction potential model. *International journal of health geographics 10*, 1 (2011), 1–16.

[87] SHAW, J. Our digital rights to the city.

[88] SINGH, P., AND LEFEVER, E. Sentiment analysis for hinglish code-mixed tweets by means of cross-lingual word embeddings. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching* (2020), pp. 45–51.

[89] SLOAN, L., AND MORGAN, J. Who tweets with their location? understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. *PloS one 10*, 11 (2015), e0142209.

[90] STATISTICS FINLAND. *Population*, 2020. Available at https://www.stat.fi/til/vrm_en.html (Accessed May 28th, 2021).

[91] STEIGER, E., WESTERHOLT, R., AND ZIPF, A. Research on social media feeds–a GIScience perspective. *European Handbook of Crowdsourced Geographic Information* (2016), 237.

[92] SUN, C., QIU, X., XU, Y., AND HUANG, X. How to fine-tune BERT for text classification? In *China National Conference on Chinese Computational Linguistics* (2019), Springer, pp. 194–206.

[93] TENNEY, I., DAS, D., AND PAVLICK, E. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019), pp. 4593–4601.

[94] THE MINISTRY OF EDUCATION AND CULTURE. *Act on the promotion of sports and physical activity*, 2015.

[95] TJONG KIM SANG, E. F., AND DE MEULDER, F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003* (2003), pp. 142–147.

[96] ULČAR, M., AND ROBNIK-ŠIKONJA, M. FinEst BERT and CroSloEngual BERT. In *International Conference on Text, Speech, and Dialogue* (2020), Springer, pp. 104–111.

[97] UNIVERSITY OF JYVÄSKYLÄ. LIPAS sport facility GIS-database. Online; Accessed: 2021-01-17.

[98] VÄISÄNEN, T., HEIKINHEIMO, V., HIIPPALA, T., AND TOIVONEN, T. Exploring human–nature interactions in national parks with social media photographs and computer vision. *Conservation Biology* (2021).

[99] VIRTANEN, A., KANERVA, J., ILO, R., LUOMA, J., LUOTOLAHTI, J., SALAKOSKI, T., GINTER, F., AND PYYSALO, S. Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076* (2019).

[100] WALLGRÜN, J. O., KARIMZADEH, M., MACEACHREN, A. M., AND PEZANOWSKI, S. GeoCorpora: building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science 32*, 1 (2018), 1–29.

[101] WANG, J., HU, Y., AND JOSEPH, K. NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages. *Transactions in GIS 24*, 3 (2020), 719–735.

[102] WANG, R.-Q., MAO, H., WANG, Y., RAE, C., AND SHAW, W. Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data. *Computers & Geosciences 111* (2018), 139–147.

[103] WEISCHEDEL, R., PALMER, M., MARCUS, M., HOVY, E., PRADHAN, S., RAMSHAW, L., XUE, N., TAYLOR, A., KAUFMAN, J., AND FRANCHINI, M. Ontonotes release 5.0. *Linguistic Data Consortium, Philadelphia, PA 23* (2013).

[104] WISE, N., AND KOHE, G. Z. Sports geography: new approaches, perspectives and directions. *Sport in Society 23*, 1 (2020), 1–10. 10.1080/17430437.2018.1555209.

[105] WISE, N., AND KOHE, G. Z. Sports geography: New approaches, perspectives and directions, 2020.

[106] WOLF, T., DEBUT, L., SANH, V., CHAUMOND, J., DELANGUE, C., MOI, A., CISTAC, P., RAULT, T., LOUF, R., AND FUNTOWICZ, M. HuggingFace's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[107] ZHAO, W. X., JIANG, J., WENG, J., HE, J., LIM, E.-P., YAN, H., AND LI, X. Comparing Twitter and traditional media using topic models. In *European conference on information retrieval* (2011), Springer, pp. 338–349.

[108] ZHENG, A., DUFF, B. R., VARGAS, P., AND YAO, M. Z. Self-presentation on social media: When self-enhancement confronts self-verification. *Journal of Interactive Advertising* (2020), 1–35.

[109] ZHU, D., ZHANG, F., WANG, S., WANG, Y., CHENG, X., HUANG, Z., AND LIU, Y. Understanding place characteristics in geographic contexts through graph convolutional neural networks. *Annals of the American Association of Geographers 110*, 2 (2020), 408–420.