Pairing Tweets with the Right Location

Esha(✉) and Osmar Zaïane(✉)

Alberta Machine Intelligence Institute, University of Alberta, Edmonton, AB, Canada
{esha1,zaiane}@ualberta.ca

Abstract. Twitter is used to provide location-relevant information and event updates. It is important to identify location-relevant tweets in order to harness location-relevant information and event updates from Twitter. However, the identification of location-relevant tweets is a challenging problem as the location names are not always explicit. Instead, mostly the location names are implicitly embedded in tweets. This research proposes a novel approach, labelled as *DigiCities*, to add geographical context to non-geo tagged tweets. The proposed approach helps in improving identification of location-relevant tweet by harnessing the location-specific information embedded in user-ids and hashtags included in tweets. Tweets relevant to eight cities were identified and used in classification experiments, and the use of *DigiCities* improved the overall classification accuracy of tweets into relevant city classes.

Keywords: Geolocation · Social media · Twitter · *DigiCities*

1 Introduction

The use of Twitter has become ubiquitous – organizations, governments, and individuals use it for various reasons (e.g., products and services promotion, information dissemination and event updates). Tweets are becoming digital footprints of users' expressions in real world and information provided by them have local relevance which can be utilized to understand what is happening in a geographical location by identifying trending topics, sentiments and emotions.

It is critical to identify the location-relevant tweets in order to learn what is happening in a geographical location [33,40]. Researchers such as Cheng et al. [6], and Lee et al. [17] have noted that geolocation detection is challenging to solve in the context of Twitter. There is limited geolocation information associated with a tweet in its metadata, and only a limited number of tweets would have correct geolocation information included in a tweet's metadata records. Graham et al. [9], for example, collected over 19 million tweets and found that only a fraction of tweets (approx. 0.7%) had geolocation information. Similarly, Lee et al. [17] noted that only 0.58% of 37 million tweets posted each day are geotagged. Both Chang et al. [5] and Inkpen et al. [13] noted that there is location-related data sparsity i.e., a very few tweets contain a specific city name. This is further complicated by the fact that users may include varying granular levels

R. Jallouli et al. (Eds.): ICDEc 2021, LNBIP 431, pp. 264–278, 2021.
https://doi.org/10.1007/978-3-030-92909-1_18

of location information when referring to a specific location [12]. For example, Cheng et al. [6] selected a random sample of one million Twitter users and found that "only 26% have listed a user location as granular as a city name (e.g., Los Angeles, CA); the rest are overly general (e.g., California), missing altogether, or [had] nonsensical location (e.g., Wonderland)" [5]. Also, metadata associated with tweets may not be complete and give reliable location information. For example, Watanabe et al. [37] noted that only 0.7% of tweets are geo-tagged and the metadata associated with posted tweets may not provide correct location information.

Consider a following scenario: John (a hypothetical Twitter user), resides in St. Paul, state capital of Minnesota, USA but his profile states Minneapolis as the location (St. Paul and Minneapolis are known as the twin cities). Currently, John is traveling to Toronto in Canada. He is sitting in a restaurant and watching a hockey game on TV played in Calgary, Canada, and tweets about it – "Just watched another win by #CalgaryFlames an amazing game played @TheSaddledome #YYC".

Based on this scenario, Calgary is actually the event-related location for this tweet, while the other two geolocations captured in the metadata record ('Minneapolis' from the Twitter profile and 'Toronto' from the posted tweet) are not relevant to the content of the posted tweet. This scenario re-iterates the argument that the location information in metadata records may not be relevant to a tweet's content. In a number of cases, a tweet content will have relevant, contextual location-related information, which can be exploited to identify appropriate location that users are referring to in their tweets. Thus, we propose an approach, labeled as *DigiCities*, to add geographical context to non-geo tagged tweets, which harnesses such information from tweets to identify location-relevant tweets. The objective of the proposed research is to enhance the identification of location relevant to tweets by utilizing information embedded in user-ids and hashtags, and tweet content. The details of the proposed approach are discussed in Sect. 3. We conducted a number of classification experiments using Weka3.6[1] to analyze the improvement in identifying locations relevant to tweets after the implementation of the proposed approach. We also evaluated whether the proposed approach can help in reducing pre-processing efforts. This was done by controlling the stopwords and stemming, the two primary pre-processing approaches used in text mining, to evaluate the overall effectiveness of our proposed approach. The findings are in Sect. 5

2 Related Work

Researchers have used different tweet features to detect locations relevant to tweets. For example, Davis et al. [8], McGee et al. [22] and Li et al. [19] investigated ways to harness the strength of social network relationships of users on Twitter to detect locations of users. Whereas, authors such as Chang et al. [5],

[1] http://www.cs.waikato.ac.nz/ml/weka/.

Cheng et al. [6] and Hong et al. [10] focused on exploiting the variations in languages and terms used by users in tweets to identify locations. Cheng et al. [6] proposed the use of probabilistic framework to detect city-level location of their users by analyzing tweet content. The authors reported that they are able to "place 51% of Twitter users within 100 miles of their actual location"(p.767). The foundation of their research work was on the idea that certain terms will be more 'local' as compared to other terms as explained by them using an example, ""howdy" which is a typical greeting word in Texas, may give the estimator a hint that the user is in or near Texas" (p.763). Similarly, Hong et al. [10] focused on harnessing term diversity due to variability in topics discussed in different geographical locations. The authors noted that users in different regions of the world might be interested in different subject content (e.g., Holi, the festival of colours in India vs. Halloween in North America), and thus, are likely to have variations in language used while discussing topics on Twitter. Such variations in language and terms found in tweet content i.e., text can be exploited to identify locations for tweets.

The identification of location relevant to tweets is further compounded by location ambiguities. These are primarily of two types: geo/geo ambiguity and geo/non-geo ambiguity [13]. An example of geo/geo ambiguity, 'Memphis' as a location name in Egypt and the US. An example of geo/non-geo ambiguity is 'Berlin' as the name of a person and also a location name in Germany. Both Paradesi [24] and Inkpen et al. [13] worked on geo/non-geo and geo/geo location disambiguation. Paradesi [24] developed a tool, Twitter tagger, which geotags the tweet content using Part of Speech tagger and Inkpen et al. [13] proposed a two-step approach to detect location and to handle location ambiguities. The authors [13] used a Conditional Random Fields (CRF) classifier using different features, including bag of words, parts of speech, adjacent token, and Gazetteer, to detect location names from tweets in the first step. They reported a number of F-scores obtained by using various combination of features at each level i.e., the city-, state- and country-level. For example: Using all features, the F-scores at token- and span-level for state and country were same at 0.85 and 0.90 respectively, and for city, the F-scores at token- and span-level were slightly different i.e., 0.83 and 0.81 respectively. Further, they developed heuristics involving a five-step disambiguation process to handle location ambiguities to further improve the location detection in the second step.

Location detection for tweets is an ongoing research issue, and newer approaches are explored to improve location detection accuracy relevant to tweets. Shen et al. [28], for example, proposed a framework labeled as NELPT, which utilises location-relevant information from three sources, location mentioned in tweet content, location included in user's profile, and location as captured at the time of posting tweet, to identify city-level appropriate location to a tweet. The authors compared their method NELPT's accuracy score with the five baseline methods accuracy scores. The authors noted that their method NELPT achieved an accuracy of 71% and the best score for one of the baselines methods was 63.3% (and the other four methods achieved scores even less than

63.3%). Singh et al. [29] work proposed the use of Markov model to identify relevant location to tweets when no specific location was mentioned in user's tweets. Their model extracted information from using the tweets posted by the user in the last 7 days and extracted the "spatio temporal sequences" from their tweets, and estimated that particular user location using a Markov model. They achieved location prediction and classification accuracy of 87% and 81% respectively (p.746).

Kumar and Singh [16] research work used Convolutional Neural Network (CNN) and extracted location-relevant terms from content of tweets to identify right locations for tweets. Thomas and Henning [31] exploit tweet's content with a number of metadata elements (e.g., user-description, user-location etc.) and proposed a neural network-based framework to predict locations for tweets. Huang et al. [12] discussed the use of a novel deep learning model for detecting tweets' location. Their model had three components that includes the use of "multi-head self-attention mechanism", originally proposed by [35] (p.4), subword features, and joint training approach involving modeling at both city- and country-level. Tian et al. [32] proposed a multi-step approach to predict Twitter user location. Their approach is "based on representation learning and label propagation (ReLP)" and it uses a number of steps to identify user location including "connection relation graph construction, user relationship filtering, user representation learning, propagation probability calculation, and user location inference" (p. 2650). Zola et al. [41] proposed an unsupervised method that used user's past tweets and Google Trends to estimate user's location. It is a multi-step approach involving collecting all the nouns from user's past tweets, and those nouns are used to calculate the Google Trend score at the city-level. These scores are used to identify city coordinates to develop synthetic spatial data for each user, which is then used to estimate user location using clustering algorithms (e.g., Gaussian Mixture Models).

Almadany et al. [2], in the location identification work, focused on detecting Twitter user's country by using a variety of publically available Twitter data related to user. They used data included in metadata records associated with user, including location, time zone, and language. In addition, they used language and location information of user's friends and followers. They collected data about users from five countries, and reported an overall accuracy score of 92.8% to detect users' location at country-level. The authors reported varying accuracies scores for each of the five countries i.e., they were able to detect users' country as Turkey with an accuracy score of 98% followed by France (96%), Spain (94%), USA (90%) and Saudi Arabia (86%). Their approach is highly dependent on the metadata information supplied by users. The claimed that they were able to detect Turkey as the country for users with higher accuracy because they feel that users from Turkey write their country name and language correctly followed by users from France, Spain and USA.

Both Ying et al. [38] and Acampora et al. [1] presented their approaches focusing on identification of geo-location for events for which information was posted on Twitter. The authors [38] used multiple data points including coor-

dinates, tweet content and geographical knowledge applied with set of rule to detect event location. Acampora et al. [1] used content of tweets to identify potential geo-location of an event using a multi-step approach. They start with clustering of tweets into event-oriented groups using the PAM (Partition Around Medoids) algorithm followed by identification of key tweets related to an event from the cluster using the OPFA (Offline Peak-Finding) algorithm. This was followed by filtering key tweets terms that do match with terms in dictionaries, and such filtered terms were considered potential location candidate names. Such terms were then checked using Google Maps API if they represent any real location name. If match found in Google Maps API, then those location names were further processed to identify the target area for an event. Their approach helped in achieving "an accuracy of about 80% by considering an error of 750 kilometers" in computing the geographic area of an event (p. 128221).

Detection of location for tweets is an ongoing research issue. We propose an unique approach to detect location relevant to tweets, and to the best of our knowledge, we have not seen the use of similar approach to detect location relevant for tweets.

3 Proposed Approach: *DigiCities*

We propose a novel approach, which creates a linkage between the digital world and the physical world. Kindberg et al. [15] noted that the information on the Internet portrays our physical world, and argued that "the physical world and the virtual world would both be richer if they were more closely linked" (p. 935). Warf and Sui [36] noted that in the age of the "metaverse", and "virtual worlds ... serve as digital equivalents to ... physical world" (p. 202). Drawing on the viewpoints of [15] and [36], a real world geographical location can be represented by multiple facets in the virtual/digital world, particularly on social media like Twitter.

The proposed approach, *DigiCities*, is the digital avatar of real world cities i.e., it is the digital identity or profile representing the real world geographical location on the web. The geographical locations are represented by facets such as People, Organizations, and Places (termed as the POP Framework). The inspiration for the POP framework came from Kindberg et al.'s [15] who divided physical entities into three key categories: people, places, and things. The term location in this research represents a geographical boundary as associated with the municipally defined boundaries for a city or town. Though geographical locations and cities are significantly different concepts, they are used interchangeably for this research. *DigiCities*, identifies members which can be categorized into three key elements of the 'POP' framework, and they are:

People: This facet represents public figures and the prominent members of a community and thus, are the face of a city. For example: City Mayor and other key people representing a given city.

Organizations: This facet represents key organizations and institutions in a city. Examples of such units include local radio channels, museums, public libraries, etc. This facet may also capture sub-units of a larger unit.

Places: This facet represents a city by its name or through the prominent spaces and landmarks. Examples of such units include legislative buildings, airport, local parks and entertainment spots.

The facets in the POP Framework i.e., people, organizations and places (POP) are also digitally reflected in tweets by handles (or user-ids, starting with '@') and hashtags (starting with '#'), these facets are semantically representing an entity i.e., a geographical location (e.g., New York) (example in Fig. 1).

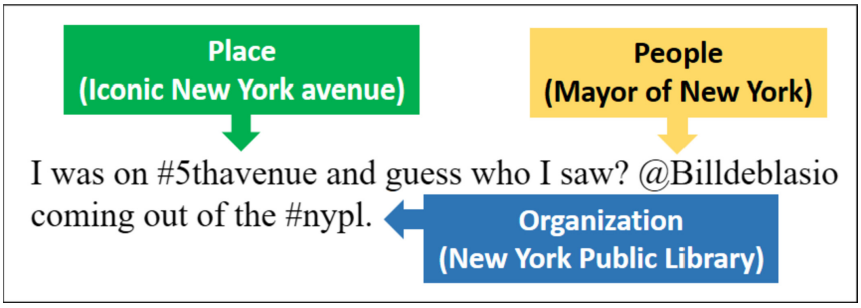

Fig. 1. Example: Tweet and the POP Framework

Such representation helps in feature convergence and/or feature strengthening [26], for example, handles and hashtags of the POP Framework are referring to (a) geographical location(s) and thereby, converging to one semantic concept i.e., a location. As noted above, data sparsity is one of the major challenges in Twitter data [13,17] and thus, has implications in the task of location detection. The feature convergence approach will help in overcoming the data sparsity issue. The elements of the POP framework have names and/or some identifying values in the physical world, and they will also embody digital names or representations in the world of Twitter. These digital names are in the form of handles (user-ids and start with '@') and hashtags (start with '#') on Twitter. The next section will provide details on *DigiCities* development, and other experimental details including dataset used in this research.

4 Methodology

It is important to note that this study used human interventions (e.g., data selection) at different stages. Normally, studies in fields like computer science focus on having huge dataset and automated processes, but studies with manual interventions, at times, are foundation to such large scale studies (e.g., validation

of results) [18]. Also, at times manual interventions are required, for example, manual coding of data [21,25,39] to create a gold standard data to evaluate experiments outcome. Thus, the methodology used in this research has limitations and are duly acknowledged but does not diminish the value of the novel framework to improve location identification relevant to tweets.

4.1 *DigiCities*: Creating Digital Profile of Cities

A total of eight urban centers in the Province of Alberta are shortlisted for this study. These geographical locations are a mix of different sized urban population centers, including the provincial capital (Edmonton), the largest city in Alberta (Calgary), a popular tourist destination (Banff), the twin-city of a larger population center (St. Albert), an industrial center (Fort McMurray), and other key cities in the Province of Alberta (Red Deer, Lethbridge, Medicine Hat). There is not much in the literature to draw upon to develop the digital profile of cities. Handles and hashtags of different People, Organizations and Places relevant to each of the above noted city are manually captured using snowball sampling technique [3] through a "recursive two-step" process.

The first step (i.e., Step 1) used the Google search engine to identify handles and hashtags of members fitting in the POP Framework. The initial seeding was done by using keywords query such as 'cityname Twitter' (e.g., Lethbridge Twitter). While, the second step (i.e., Step 2) used Google search results to connect with specific user's Twitter account, and the next set of handles were selected based on Twitter's recommendation of other handles under the 'You may also like'. Each handle was reviewed for relevancy to a city and was collected, if relevant. This process continued until the recommended handles either started repeating themselves or are no longer relevant to the city. Further, during the digital profile development of cities and data collection, it was observed that a number of handles have equivalent hashtags. For example,'@banff' and '@calgarystampede' has an equivalent hashtag of '#banff'and '#calgarystampede' respectively. Thus, all the handles were converted into equivalent hashtags to capture such occurrences. Further, a number of handles or hashtags relevant to a city had city name and its variant (e.g. MedicineHat or mhat for Medicine Hat) or airport code (if there was any) included either as prefix or suffix. Such additional digital profile terms involving the city name and airport code (note: St. Albert and Banff do not have an airport), were captured by using regular expressions (e.g., calgary in @calgarytoday, and Calgary city airport code 'yyc' in #yyctraffic). It is important to note that multi-term city names were combined into one term (e.g., 'Red Deer' into 'reddeer'). Thus, the total number of handles and hashtags, and their variants included in each city's profile were (count in bracket): Banff (114), Calgary (214), Edmonton (198), Fort McMurray (100), Lethbridge (98), Medicine Hat (46), Red Deer (112) and St. Albert (72).

4.2 Append Strategy and Replace Strategy

Replace Strategy and Append Strategy are applied to converge and strengthen features in tweets where a location is represented semantically through various facets of the POP Framework. The append strategy implementation led to the inclusion of the city name (e.g., Reddeer) in tweets when the terms of tweets matched with the terms in the digital profile of a city. The replace strategy implementation led to the replacement of terms in tweets by the city name when the terms of tweets matched with the terms in the digital profile of a city. Table 1 provides an example of a tweet relevant to New York (Original Tweet in Table 1). The terms '#LGA', '@Broadwaycom', and '#biggapple' in the example tweet matched with the terms in the city of New York profile. In the append strategy, city name, 'NewYork', is appended after the matching terms, #LGA, @Broadwaycom, and #biggapple (Append Strategy in Table 1). In replace strategy, the matching terms, #LGA, @Broadwaycom, and #biggapple are replaced by the city name, i.e., 'NewYork' (Replace Strategy in Table 1).

Table 1. Example of implementation of replace strategy and append strategy

Original Tweet	Just landed at #LGA and went straight to @Broadwaycom so see #Aladdin. This is why I love the #biggapple
Append Strategy	Just landed at #LGA **newyork** and went straight to @Broadwaycom **newyork** so see #Aladdin. This is why I love the #biggapple **newyork**
Replace Strategy	Just landed at **newyork** and went straight to **newyork** so see #Aladdin. This is why I love the **newyork**

4.3 Data, Experimentation, Algorithms and Evaluation

Twitter data was collected intermittently for approximately for 12 months, January 12, 2017 to December 30, 2017, using a dedicated API [27]. The initial corpus had over 700,000 tweets related to the Province of Alberta in Canada [27]. It was a purposeful shortlisting of tweets. The selection of tweets was terminated once the tweet count reached 500 for a city. The purposeful criteria included selecting a tweet if it was in the English language, and the coder was able to assess tweet's relevancy to a specific city (as discussed in the scenario in Introduction Section). There were varying numbers of tweets for each of the eight cities and purposefully 500 tweets were manually selected for each city [30] plus 500 random tweets were selected for one additional category of 'Others' to capture tweets not belonging to any city class. Thus, a total of 4,500 tweets were used in this research and it was deemed as an appropriate number of tweets considering they were manually reviewed and selected by one coder. Authors like Rogstad [25] used 1,500 tweets and noted that "[t]his was considered a manageable number of tweets for manual coding" (p.146) as it is costly both in terms of "time and effort" ([23] p.1230).

Only basic data cleaning was done and both stopwords removal and stemming was not done at this stage. Basic cleaning includes removal of URLs, special characters, and white spaces between handle (@) (or hashtag (#)) symbol,

and the term following it (e.g., '@' was joined with the adjacent term). This is labelled as No Preprocessing in Table 2 and No_Pre in Fig. 2. The original tweets formed the *Baseline Data*. The data created after implementing the append and replace strategies created the *Append Data* and the *Replace Data* respectively (see Table 2 for example). Preprocessing (e.g., stemming and stopwords removal) is critical in text mining, and depending on data quality, it can be a time consuming activity. In this research, the two preprocessing procedures i.e., removal of stopwords and stemming impact were evaluated in combination with the *DigiCities* to investigate if the proposed approach of *DigiCities* can help in reducing preprocessing steps. Thus, each data type i.e., Baseline Data, Append Data and Replace Data, had three variants of data. For example: Baseline data had the following variants: Baseline Data (original dataset), Baseline Data after removing stopwords, and Baseline Data after stemming.

The classification experiments were done using three well-known algorithms, Naïve Bayes, kNN (k = 3) & SMO (a SVM variant), as implemented in Weka3.6, to evaluate the effectiveness of our proposed approach. Previous research work in the classification area suggests that all the three classification algorithms can achieve good results in text classification [4,14]. Five fold cross validation was performed, and the authors, such as Hsu et al. [11], Cho et al. [7], and Mahajan et al. [20]), suggested that the cross validation (e.g.,. five-fold) can help in mitigating the issue of overfitting. The results were evaluated using standard evaluation measures including precision and recall, and accuracy. Accuracy was defined as the total number of tweets correctly classified into their respective classes divided by the total number of tweets (i.e., 4,500). A total of 27 classification experiments were conducted (i.e., nine data variants x three algorithms).

5 Findings

A total of 3,780 terms matched with the terms in eight city profiles in 4,000 tweets. The number of terms matching varied for cities. Two cities, Banff and Red Deer had lower number of terms matching at 340 and 341 respectively. While, cities like Calgary, Edmonton, Fort McMurray and Lethbridge had over 500 matches and Lethbridge had 553 matches, highest among all cities. The matching of profile terms with terms in tweets is dependent on both the number of hashtags and user handles in tweets, and the number of terms in city's digital profile. The following sub-sections will discuss the impact of our proposed approach, *D*igiCities, on the classification of tweets into appropriate city-based classes.

5.1 Impact of *DigiCities* (Prior to Preprocessing)

The classification experiments results on the baseline data showed improvement in the classification accuracy scores after the implementation of our proposed approach *DigiCities* over the baseline data (i.e., before implementation of *DigiCities* approach). First, the accuracy scores of all the three algorithms had significantly improved over baseline data. kNN results on baseline data yields the

lowest accuracy score of 47.6% followed by NB with 69.9% and SMO with 87.8%. After the implementation of our approach, irrespective of the use of replace or append strategy, the accuracy scores improved for each algorithm over their respective baseline accuracy scores. The accuracy score for kNN improved from 47.6% to 56.1% and 69.9% with replace and append strategy respectively. These scores are statistically significantly different as demonstrated by the chi-square test as the p-value (1.40E−98) is less than 5%.

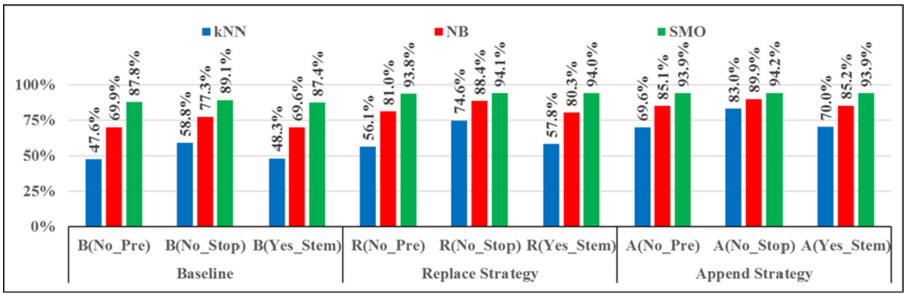

Fig. 2. Accuracy scores

Similarly, for both NB and SMO, the accuracy score improved with the implementation of our approach. In the case of NB, the accuracy improved from 69.9% to 81.0% and 85.1% with the use of replace and append strategy respectively. The p-value from chi-square test (9.86E-73) is less than 5% showing that the scores are statistically significantly different. While for SMO, the accuracy score improved from 87.8% to 93.8% and 93.9% with the use of replace and append strategy respectively. The chi-square test yielded the same result i.e., these scores are statistically significantly different as p-value (1.36E-32) was less than 5%. Among all the three algorithms, the highest relative improvement is observed in the kNN accuracy score followed by NB, and lowest was for the SMO.

Both the precision and recall improved for all the three algorithms after the implementation of append or replace strategies over the baseline weighted average precision and recall (see Table 2). For example, the precision for kNN with the baseline was 0.66 and it changed to 0.75 and 0.72 for the append and replace strategies respectively (Table 2). Interestingly, the append strategy gave relatively better precision and recall as compared to the replace strategy for both kNN and NB algorithms. While for the SMO algorithm, the precision and recall achieved was almost the same with the use of the append and replace strategies.

5.2 Impact of *DigiCities* and Preprocessing

It is noted that the stopwords removal and stemming can help in improving classification accuracy of text data [34] but these are additional steps which has to be followed to achieve a good outcome. This research aimed at evaluating if our proposed approach can help in reducing any preprocessing on tweet data. The focus of preprocessing was on two facets, stopwords removal and stemming.

Table 2. Precision and recall scores

Algorithms	Measures	No Preprocessing			Stopwords Removed			Stemming Applied		
		Baseline	Append	Replace	Baseline	Append	Replace	Baseline	Append	Replace
kNN	Precision	0.66	0.75	0.72	0.68	0.86	0.83	0.65	0.75	0.71
	Recall	0.48	0.70	0.56	0.59	0.83	0.75	0.48	0.70	0.58
NB	Precision	0.75	0.88	0.84	0.82	0.89	0.92	0.75	0.88	0.83
	Recall	0.70	0.85	0.81	0.77	0.90	0.88	0.70	0.80	0.85
SMO	Precision	0.91	0.95	0.95	0.93	0.96	0.95	0.90	0.95	0.95
	Recall	0.88	0.94	0.94	0.89	0.94	0.94	0.87	0.94	0.94

Stopwords Removal: Comparing the accuracy scores in Fig. 2, and precision and recall score in Table 2 show that the stopwords removal (labelled as No_Stop Fig. 2) had a varying level of positive impact on the accuracy scores for all the three algorithms. As expected, the accuracy scores improved significantly by removing stopwords and after the use of our proposed approach of *DigiCities* for both kNN (e.g., 47.6% for B(No_Pre) to 83% for A(No_Stop) and NB (e.g., 69.9% for B(No_Pre) to 89.9% for A(No_Stop). Interestingly, for the SMO algorithm, the removal of stopwords and without implementing our *DigiCities* approach, the change in the accuracy score was marginal i.e., from 87.8% for B(No_Pre) to 89.1% for B(No_Stop). However, after implementing our strategy *DigiCities*, the score changed from 93.9% for A(No_Pre) to 94.2% for A(No_Stop). Following the removal of stopwords, both the weighted average precision and recall improved after the implementation of our strategies (Table 2).

Stemming Applied: Results in Fig. 2 show that after implementing stemming (labelled as Yes_Stem in Fig. 2), the impact on the accuracy scores was only marginal for all three algorithms as compared to the impact after removal of stopwords. The results also show that accuracy scores improved after stemming with the implementation of *DigiCities* can only be attributed to our proposed approach. Following the implementation of stemming, both the weighted average precision and recall improved after the implementation of the append strategy and the replace strategy over the baseline precision and recall (Table 2).

5.3 Append Strategy Vs. Replace Strategy

Both append and replace strategies helped in improving the classification accuracy of all the three algorithms (Fig. 2). SMO achieved the highest accuracy score, and the improvement was by 6% over the baseline score for any data variants, and the gain made by the use of append or replace strategy was nearly the same. Also, there was no statistical difference in the accuracy scores achieved by the use of append or replace strategy using SMO (p value: 0.9).

kNN had the highest increase in the accuracy score (by 22%) followed by NB (by 15%) with the use of append strategy as compared to gain using replace strategy, the gain was relatively less for both kNN (8.5%) and NB (11.1%) over the baseline accuracy scores (Fig. 2). Further, based on (Fig. 2), the chi-square tests

results reveal that there is statistical difference in the accuracy scores between the append strategy and the replace strategy for kNN (p value: 5.25E−40) and NB (p value: 2.01E−07).

The key findings includes: a) *DigiCities* can help in improving the classification accuracy score by using either append or replace strategy; b) SMO algorithm in general proved to be the better choice among the three algorithms; c) With the use of our approach and SMO, both removal of stopwords and stemming may not play a critical role; d) Removal of stopwords with our proposed approach of *DigiCities* will positively impact classification accuracy for both kNN and NB algorithms; e) Stemming on tweet data may not play a critical role, particularly when used with our approach; f) The append strategy is better as compared to the replace strategy to implement *DigiCities* when using kNN and NB algorithms but with SMO either strategy would work.

6 Conclusion, Limitations and Future Work

The accuracy scores for all the three algorithms, kNN, NB and SMO, improved after the implementation of the *DigiCities* approach and this suggests that *DigiCities* can help in identifying location-relevant tweets by harnessing city-relevant information from tweet content such as hashtags and handles. Further, among both the strategies, the append strategy gave relatively better classification accuracy score over the replace strategy. Among the three algorithms, the SMO algorithm performance was best as compared to kNN and NB algorithms.

The study has a number of limitations. For example, the study includes only eight cities from the Province of Alberta. The proposed approach, *DigiCities*, needs to be tested further by including more cities from other regions of Canada and other countries. There is potential of researcher's bias in data preparation as tweets for different cities were manually selected. The digital profile of cities were manually created and there is room to make them more comprehensive. Further, a number of times hashtags and handles used in tweets do not categorize into any of the existing element of the POP framework and thus such hashtags and handles are not included in the digital profile of a city, then in such cases, the city relevant features in a tweet will not get strengthened.

The use of the proposed approach of *DigiCities* has improved the overall accuracy as well as the precision and recall. We plan to extend this work in multiple ways and aim to address some of the limitations in future work. First, we aim to test the proposed approach by increasing both the diversity of cities and the size of dataset. Second, we aim to develop an automated process to establish more comprehensive digital profiles by web scraping of Twitter pages on the basis of city's geographical data. Third, we aim to extend the POP Framework by adding new facets such as local language and seasonal terms e.g., hashtags or user-ids of yearly occurring events in a city like Mardi Gras Carnaval in New Orleans, USA. Fourth, we aim to test our approach using other classification algorithms and examine the impact resulting from varying of hyperparameters. Finally, we aim to implement this approach in combination with other approaches (e.g., Inkpen et al. [13]) to make improvements in location detection and disambiguation.

References

1. Acampora, G., Anastasio, P., Risi, M., Tortora, G., Vitiello, A.: Automatic event geo-location in Twitter. IEEE Access **8**, 128213–128223 (2020)
2. Almadany, Y., Saffer, K.M., Jameil, A.K., Albawi, S.: A novel algorithm for estimation of Twitter users location using public available information. Int. J. Smart Sens. Intell. Syst. **13**(1), 1–10 (2020)
3. Biernacki, P., Waldorf, D.: Snowball sampling: problems and techniques of chain referral sampling. Sociol. Methods Res. **10**(2), 141–163 (1981)
4. Bijalwan, V., Kumar, V., Kumari, P., Pascual, J.: KNN based machine learning approach for text and document mining. Int. J. Database Theory Appl. **7**(1), 61–70 (2014)
5. Chang, H.w., Lee, D., Eltaher, M., Lee, J.: @Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage. In: IEEE International Conference on Advances in Social Networks Analysis and Mining, pp. 111–118 (2012)
6. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geo-locating Twitter users. In: ACM International Conference on Information and Knowledge Management, pp. 759–768 (2010)
7. Cho, H.h., Lee, S.h., Kim, J., Park, H.: Classification of the glioma grading using radiomics analysis. PeerJ **6**, e5982 (2018)
8. Davis Jr, C.A., Pappa, G.L., de Oliveira, D.R.R., de L. Arcanjo, F.: Inferring the location of Twitter messages based on user relationships. Trans. GIS **15**(6), 735–751 (2011)
9. Graham, M., Hale, S.A., Gaffney, D.: Where in the world are you? Geolocation and language identification in Twitter. Prof. Geogr. **66**(4), 568–578 (2014)
10. Hong, L., Ahmed, A., Gurumurthy, S., Smola, A.J., Tsioutsiouliklis, K.: Discovering geographical topics in the Twitter stream. In: International Conference on World Wide Web, pp. 769–778. ACM (2012)
11. Hsu, C.W., Chang, C.C., Lin, C.J., et al.: A practical guide to support vector classification (2003)
12. Huang, C.Y., Tong, H., He, J., Maciejewski, R.: Location prediction for tweets. Front. Big Data **2**, 5 (2019)
13. Inkpen, D., Liu, J., Farzindar, A., Kazemi, F., Ghazi, D.: Detecting and disambiguating locations mentioned in Twitter messages. In: Gelbukh, A. (ed.) CICLing 2015. LNCS, vol. 9042, pp. 321–332. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18117-2_24
14. Joachims, T.: Making large-scale SVM learning practical. Technical report, SFB 475: Komplexitätsreduktion in Multivariaten ...(1998)
15. Kindberg, T., et al.: People, places, things: web presence for the real world. Mobile Netw. Appl. **7**(5), 365–376 (2002)
16. Kumar, A., Singh, J.P.: Location reference identification from tweets during emergencies: a deep learning approach. Int. J. Disaster Risk Reduction **33**, 365–375 (2019)
17. Lee, K., Ganti, R.K., Srivatsa, M., Liu, L.: When Twitter meets foursquare: tweet location prediction using foursquare. In: International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, pp. 198–207 (2014)
18. Leon, A.C., Davis, L.L., Kraemer, H.C.: The role and interpretation of pilot studies in clinical research. J. Psychiatr. Res. **45**(5), 626–629 (2011)

19. Li, R., Wang, S., Chang, K.C.C.: Multiple location profiling for users and relationships from social network and content. VLDB **5**(11), 1603–1614 (2012)
20. Mahajan, R., Viangteeravat, T., Akbilgic, O.: Improved detection of congestive heart failure via probabilistic symbolic pattern recognition and heart rate variability metrics. Int. J. Med. Inform. **108**, 55–63 (2017)
21. Massey, P.M., Leader, A., Yom-Tov, E., Budenz, A., Fisher, K., Klassen, A.C.: Applying multiple data collection tools to quantify human papillomavirus vaccine communication on Twitter. J. Med. Internet Res. **18**(12), e318 (2016)
22. McGee, J., Caverlee, J., Cheng, Z.: Location prediction in social media based on tie strength. In: International Conference on Information & Knowledge Management, pp. 459–468. ACM (2013)
23. Ogan, C., Varol, O.: What is gained and what is left to be done when content analysis is added to network analysis in the study of a social movement: Twitter use during Gezi park. Inf. Commun. Soc. **20**(8), 1220–1238 (2017)
24. Paradesi, S.M.: Geotagging tweets using their content. In: Twenty-Fourth International FLAIRS Conference (2011)
25. Rogstad, I.: Is Twitter just rehashing? Intermedia agenda setting between Twitter and mainstream media. J. Inf. Technol. Polit. **13**(2), 142–158 (2016)
26. Saif, H., He, Y., Alani, H.: Semantic sentiment analysis of Twitter. In: Cudré-Mauroux, P., et al. (eds.) ISWC 2012. LNCS, vol. 7649, pp. 508–524. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35176-1_32
27. Samuel, H., Zaïane, O., Martz, P.: Supporting digital epidemiology in Alberta via Twitter tracking. In: International Conference on Biomedical and Health Informatics (2017)
28. Shen, W., Liu, Y., Wang, J.: Predicting named entity location using Twitter. In: IEEE International Conference on Data Engineering (ICDE), pp. 161–172 (2018)
29. Singh, J.P., Dwivedi, Y.K., Rana, N.P., Kumar, A., Kapoor, K.K.: Event classification and location prediction from tweets during disasters. Ann. Oper. Res. **283**(1), 737–757 (2019)
30. Teddlie, C., Yu, F.: Mixed methods sampling: a typology with examples. J. Mixed Methods Res. **1**(1), 77–100 (2007)
31. Thomas, P., Hennig, L.: Twitter geolocation prediction using neural networks. In: Rehm, G., Declerck, T. (eds.) GSCL 2017. LNCS (LNAI), vol. 10713, pp. 248–255. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73706-5_21
32. Tian, H., Zhang, M., Luo, X., Liu, F., Qiao, Y.: Twitter user location inference based on representation learning and label propagation. In: Proceedings of the Web Conference 2020, pp. 2648–2654 (2020)
33. Tsou, M.H.: Mapping cyberspace: tracking the spread of ideas on the internet. In: International Cartographic Conference (2011)
34. Uysal, A.K., Gunal, S.: The impact of preprocessing on text classification. Inf. Process. Manag. **50**(1), 104–112 (2014)
35. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017). https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
36. Warf, B., Sui, D.: From GIS to neogeography: ontological implications and theories of truth. Ann. GIS **16**(4), 197–209 (2010)
37. Watanabe, K., Ochi, M., Okabe, M., Onai, R.: Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In: International Conference on Information and Knowledge Management, pp. 2541–2544. ACM (2011)

38. Ying, Y., Peng, C., Dong, C., Li, Y., Feng, Y.: Inferring event geolocation based on Twitter. In: Proceedings of the 10th International Conference on Internet Multimedia Computing and Service, pp. 1–5 (2018)
39. Zahra, K., Imran, M., Ostermann, F.O.: Automatic identification of eyewitness messages on Twitter during disasters. Inf. Process. Manag. **57**(1), 102107 (2020)
40. Zheng, X., Han, J., Sun, A.: A survey of location prediction on Twitter. IEEE Trans. Knowl. Data Eng. **30**(9), 1652–1671 (2018)
41. Zola, P., Ragno, C., Cortez, P.: A google trends spatial clustering approach for a worldwide Twitter user geolocation. Inf. Process. Manag. **57**(6), 102312 (2020)