Towards Real-Time, Country-Level Location Classification of Worldwide Tweets

Arkaitz Zubiaga\textsuperscript{1}, Alex Voss\textsuperscript{2}, Rob Procter\textsuperscript{1}, Maria Liakata\textsuperscript{1}, Bo Wang\textsuperscript{1}, Adam Tsakalidis\textsuperscript{1}

\textsuperscript{1} University of Warwick
Coventry, UK
\textsuperscript{2} University of St Andrews
St Andrews, UK

a.zubiaga@warwick.ac.uk

\section{Introduction}

Social media are increasingly being used in the scientific community as a data source that enables the study of multiple aspects of society, ranging from the computational development of tools that make use of knowledge garnered from social media in real time, to analytical studies in the social sciences. Thanks to the availability of a public API that enables free collection of a significant amount of data, Twitter has become a key tool for a diverse range of scientific research [49].

Having Twitter as a new kind of data source, researchers have looked into the development of tools for real-time trend analytics [32], [54] or early detection of newsworthy events [46], as well as into analytical approaches for understanding the sentiment expressed by users towards a target [23], [26], [48] or public opinion on a specific topic [3]. However, Twitter data lacks reliable demographic details of users that would enable a representative sample to be collected and/or a focus on a specific user subgroup [35]. One of these missing demographic details is a user’s country of origin. The only option then for the researcher is to try to infer such demographic characteristics before attempting the intended analysis.

This has motivated a growing body of research in recent years looking at different ways of determining automatically the user’s country of origin and/or – as a proxy for the former – the location from which tweets have been posted [1]. Most of the previous research in inferring tweet geolocation has tried to classify tweets by state or by city within a specific geographical area or country (e.g. the USA). The few cases that have dealt with a global collection of tweets used an extensive set of features that cannot realistically be extracted in a real-time, streaming context (e.g. user tweeting history and/or social networks). An ability to classify tweets by location in real time is crucial for applications exploiting social media updates as social sensors that enable tracking topics and learning about emerging events and breaking news. To the best of our knowledge, our work is the first to deal with global tweets in any language, using only those features present within the content of a tweet and its associated metadata. Moreover, our experiments complement previous work by investigating the extent to which a classifier trained on historical tweets can still be used effectively on newly harvested tweets.

Motivated by the need to develop an application that can identify the trending topics within a specific country at a given point in time\textsuperscript{1}, here we document the development of a classifier that can geolocate tweets by country of origin in real-time. Given that within this scenario it is not feasible to collect additional data to that readily available from the Twitter stream, we explore the usefulness of eight

\textsuperscript{1} http://www.bbc.co.uk/programmes/b04p59vr
different features inherent in each tweet for determining its geolocation. All of the eight features we consider for the classification of tweets by country are readily available from a tweet object as retrieved from the Twitter API. We perform classification using each of the features alone, but also in feature combinations. We explore the ability to perform the classification on as many as 217 countries, or in a reduced subset of the top 25 countries, as judged by tweet volume. The use of two datasets, collected in October 2014 and October 2015, gives additional insight into whether historical Twitter data can be used to classify new instances of tweets. These two datasets, consisting of more than 5 million tweets with associated country codes, are being made publicly available, thus enabling replication and further research.

Our methodology enables us to perform a thorough analysis of tweet geolocation, revealing insights into the best approaches for an accurate country-level location classifier for tweets. Our experiments and analysis shed light on how to best build an application that classifies tweets by country accurately in real time, whether the goal is to organise content by country or whether one wants to identify content from a specific country, filtering out the rest. While the content of a tweet is one of the most commonly used features in previous work, we observe that the validity of historical content fades over time, and other metadata associated with the tweet, such as the language of the tweet, the name of the user, or the time zone in which the user is located, are more appropriate features that lead to more accurate classification. We also perform a per-country analysis for the top 25 countries in terms of tweet volume, exploring how different features lead to optimal classification for different countries, as well as discussing limitations when dealing with some of the most challenging countries.

2 Related Work

A growing body of research deals with the automated inference of demographic details of Twitter users [35]. Researchers have attempted to infer attributes of Twitter users such as age [22], gender [5], income [40] and socioeconomic status [28]. Work by Huang et al. [21] has also tried to infer the nationality of users; this work is different from ours in that the country where the tweets were posted from, was already known.

What motivates the present study is the increasing interest in inferring the geographical location of either tweets or Twitter users [1]. The automated inference of tweet location has been studied for different purposes ranging from data journalism [19] to public health [14]. As well as numerous different techniques, researchers have relied on different settings and pursued different objectives when conducting experiments. Table 1 shows a summary of the previous work reported in the scientific literature, outlining the features that each study used to classify tweets by location, the geographic scope of the study, the languages they dealt with, the classification granularity they tried to achieve and used for evaluation, and whether single tweets, aggregated multiple tweets and/or user history were used to train the classifier.

Most of the previous studies on automated geolocation of tweets has assumed that the tweet stream includes only tweets from a specific country. The majority of these studies have focused on the United States, classifying tweets either at a city or state level. One of the earliest studies is that by Cheng et al. [7], who introduced a probabilistic, content-based approach that identifies the most representative words of each of the major cities in the USA; these words are then used to classify new tweets. They incorporate different techniques to filter words, such as local and state-level filtering, classifying up to 51% of Twitter users accurately within a 100 mile radius. Their approach, however, relies on making use of the complete history of a user, and was tested only for users with at least 1,000 tweets in their timeline.

Most of the other studies documented in the literature have also relied on tweet content, using different techniques such as topic modelling to find locally relevant keywords that reveal a user’s likely location [7], [6], [44], [17], [30], [33], [8], [27]. Another widely used technique relies on the social network that a user is connected to, so as to infer a user’s location from that of their followers and followees [25], [44], [45]. While the approaches summarised will work well for certain applications, retrieving the tweet history for each user or the profile information of all of a user’s followers and followees is not feasible in a real-time scenario. Hence, in this context, a classifier needs to deal with the additional challenge of having to rely only on the information that can be extracted from a single tweet.

Only a handful of studies have relied solely on the content of a single tweet to infer its location [13], [51], [2], [13], [16]. Again, most of these have actually worked on very restricted geographical areas, with tweets being limited to different regions, such as the United States in two of them [15], [51], four different cities in another work [16], and New York only in the other case [13]. Bo et al. [2] did focus on a broader geographical area, including 3.7k cities all over the world. Nevertheless, their study focused on a limited number of cities, disregarding other locations, and only classified tweets written in English.

When it comes to geolocation classification granularity, the majority of studies have aimed at city-level classification. While this provides fine-grained classification of tweets, it also means that a limited number of cities can be considered, ignoring other cities and towns. Only Han et al. [17] performed country-level classification, although they also restricted their study to tweets posted from a limited number of cities. This basically means that tweets posted from cities other than the ones under consideration are removed from the stream, which poses an additional challenge for the classifier. In our study, we take as input the stream of tweets with content originating from any country, to classify, at the country-level, each tweet according to its origin. Our work hence considers any tweet retrieved from a stream, without excluding any locations or languages.

In summary, as far as we are aware, no previous work has dealt with the multiple features available within a tweet, as retrieved from the Twitter streaming API, to determine the location of any tweet posted from anywhere in the
characteristics of previous studies of automated geolocation of tweets or Twitter users. The present study, in the last row, represents the first attempt to deal with global tweets and in any language by using only features that are readily available within the body of a tweet or its metadata.

3 Datasets

For training our classifier, we rely on the most widely adopted approach for the collection of a Twitter dataset with tweets categorised by location. This involves using the Twitter API endpoint that returns a stream of geolocated tweets posted from within one or more specified geographic bounding boxes. In our study, we set this bounding box to be the whole world (i.e., [-180,-90,180,90]) in order to retrieve tweets worldwide. This way, we collected streams of global geolocated tweets for two different week long periods: 4-11 October, 2014 (TC2014) and 22-28 October, 2015 (TC2015). This led to a collection of 4,155,763 geolocated tweets in TC2014 and 897,341 geolocated tweets in TC2015.

Having these tweets geolocated with the specific coordinates of the user’s location, we then inferred the name of that location. We used Nominatim for this, whose reverse geocoding feature enabled us to retrieve detailed information of the location pointed by the coordinates given as input. From Nominatim’s output, we made use of the country code in our experiments that aimed at country level classification of tweets. As a result, we had all the tweets in TC2014 and TC2015 categorised by country, which we then used as the ground truth for our classification experiments.

The more than 5 million tweets in these two datasets are categorised into 217 different countries. It is worthwhile mentioning that, as one would expect, the resulting datasets are clearly imbalanced, where only a few countries account for most of the tweets. The first country by number of tweets is the United States (20.99%), followed by Indonesia (14.01%) and Turkey (8.50%). The 10 most prominent countries on Twitter in our datasets account for 72.98% of the tweets, while the 25 most prominent countries account for 90.22%. Figure 1 shows a heat map of popularity by country in our datasets.

The resulting datasets, both TC2014 and TC2015, are publicly available.

4 Country-Level Location Classification for Tweets

In this study, we define the country-level location classification task as that in which, given a single tweet as input, a classifier has to determine the country of origin of the tweet. We argue for the sole use of the content and metadata provided in a single tweet, which are accessible in a scenario where one wants to classify tweets by country in the tweet stream and in real-time. Most existing approaches have looked at the history of a Twitter user or the social

Authors | Features | Geographic scope | Languages | Classif. granularity | Tweets/Users
--- | --- | --- | --- | --- | ---
Eisenstein et al. [15] | Tweet content | US only | All | Grid cells | Tweets
Cheng et al. [17] | Tweet content | US only | All | City-level | Users
Wing and Baldridge [51] | Tweet content | US only | All | Grid cells | Tweets
Bo et al. [2] | Tweet content | Worldwide, 3.7k cities | English | City-level | Tweets
Chen et al. [6] | Tweet content | Worldwide | English | City-level | Users
Jurgens [25] | Social network | Worldwide | All | City-level | Users
Rodrigues et al. [44] | Tweet content + social network | Brazil only, 3 cities | Portuguese | City-level | Users
Kout et al. [48] | Social network | UK only | English | City-level | Users
Doran et al. [13] | Tweet content | New York only | English | Grid cells | Tweets
Graham et al. [16] | Tweet content | 4 metropolitan areas | 9 languages | City-level | Tweets
Han et al. [17] | Tweet content | Worldwide, 3.1k cities | English | City and country | Users
Lee et al. [29] | Tweet content | Manhattan only | English | Fine-grained location | Users
Mahmud et al. [33] | Tweet content + user activity | US only | English | City-level | Users
Compton et al. [8] | Social network | Worldwide | All | City-level | Users
Krishnamury et al. [27] | Tweet content | US only | All | City-level | Users

Present work | 8 tweet features | Worldwide | All | Country-level | Tweets

Table 1

Characteristics of previous studies of automated geolocation of tweets or Twitter users. The present study, in the last row, represents the first attempt to deal with global tweets and in any language by using only features that are readily available within the body of a tweet or its metadata.

2. Twitter API’s ‘statuses/filter’ endpoint: https://dev.twitter.com/streaming/reference/post/statuses/filter

3. http://wiki.openstreetmap.org/wiki/Nominatim

4. Datasets, as well as details enabling reproducibility, are available through figshare: https://figshare.com/articles/Tweet_geolocation_5m/3168529
network derivable from a user’s followers and followees, which would not be feasible in our real-time scenario.

4.1 Classification Techniques

We have used Support Vector Machines (SVMs) [11], [24] to perform classification experiments using different sets of features available in a tweet, and to assess the usefulness of each of the features for the location classification task. Other classifiers such as random forests, naïve bayes and logistic regression were also tested in initial experiments, but SVMs were empirically found to perform best. SVMs are a supervised learning algorithm that defines the maximum margin hyperplane that separates the different classes/locations in the instance space of the training set. Moreover, using SVMs it is also possible to determine the confidence of a prediction, in terms of the distance between the data point and the hyperplane, also known as the margin. Originally, SVMs were defined as binary classifiers that attempted to satisfy the optimisation function in Equation 1 [4], [11].

\[
\min \left[ \frac{1}{2} ||w||^2 + C \sum_{i=1}^{l} \xi_i \right] 
\]

Subject to:

\[
y_i (w \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0
\]

where \( C \) is the penalty parameter, \( \xi_i \) is a stack variable for the \( i^{th} \) element, \( l \) is the number of labelled instances, and \( d \) is the sigma parameter that defines the non-linear mapping from the input space to some high-dimensional feature space.

While SVMs were originally limited to binary classification, different alternatives now exist for multiclass classification [20], including native multiclass implementations and the use of multiple binary classifiers for pairwise classifications that are ultimately combined to make the final, multiclass decision. The best-known approaches for combining binary classifiers to produce a multiclass output include “one-against-all”, where a binary classifier is created for each category with respect to the rest, and “one-against-one”, where a binary classifier is created for each pair of categories. After empirical experiments, and given that the number of binary classifiers increases substantially for multiclass problems with many categories, we opted for using svm-multiclass\(^5\) as an implementation of native multiclass SVMs [12], [50]. Different from the classical “one-against-all” and “one-against-one” approaches, a native multiclass SVM casts the multiclass problem as a single constrained optimization problem. Modifying the optimisation function for binary SVMs, multiclass SVMs for a task with \( k \) categories attempt to optimise the function in Equation 2 [50].

\[
\min \left[ \frac{1}{2} \sum_{m=1}^{k} ||w_m||^2 + C \sum_{i=1}^{l} \sum_{m \neq y_i} \xi_{mi} \right] 
\]

Subject to:

\[
w_{y_i} \cdot x_i + b_{y_i} \geq w_m \cdot x_i + b_m + 2 - \xi_{mi}, \quad \xi_{mi} \geq 0
\]

We built separate SVM classifiers for each of the features inherent to a tweet. This enabled us to evaluate the extent

5. http://svmlight.joachims.org/svm_multiclass.html
to which each of the features is useful on its own for the classification task. To test the effectiveness of using multiple features for the classification, we aggregated the output of the classifiers for the features being combined. With SVMs, this can be achieved by adding up the confidence values, i.e., the margin, output by each of the classifiers. A multiclass SVM classifier outputs a margin for each of the categories, which can be positive or negative. When combining the output of multiple classifiers for a single data point, their margins on each of the categories are added up and, ultimately, the category with the largest aggregated margin is selected as the predicted category. This approach of combining the output of multiple SVM classifiers is known as SVM committees [47], [52]. An alternative approach that combined features together in a single SVM classifier was tested, but we found that some features were overly dominating over others, given the different types of features we deal with.

As a baseline approach, commonly used in the literature, we made use of the GeoNames geographical database6. The user location is a string optionally specified by users in their profile settings, and is made available in the tweet’s metadata. It can be used here as input to the GeoNames database, which will return a likely location translated from that string. GeoNames provides a list of the most likely locations for a given string, based on either relevance or population, from which we took the first element. While GeoNames can be very effective for certain location names that are easy to map, the use of this feature is limited to users who opt to specify a non-empty location string in their settings (67.1% in our datasets), and will fail with users whose location is not a valid country or city name (e.g., somewhere in the world). The location specified in the user’s profile has been used before to infer a user’s location, although it is known to lead to low recall [36]. Here, we used this approach, using a database to translate user locations as a baseline, and explored whether, how, and the extent to which, a classifier can outperform it. For this baseline approach, we query GeoNames with the location string specified by the user, and pick the first option output by the service. To make a fairer comparison with our classifiers, since GeoNames will not be able to determine the location for users with an empty location field, we default GeoNames’ prediction for those tweets to be the majority country, i.e., the United States.

4.2 Experiment Settings
Within the TC2014 dataset, we created 10 different random distributions of the tweets for cross-validation, each having 50% of the tweets for training, 25% for development and 25% for testing. The performance of the 10 runs on the test set were ultimately averaged to get the final performance value. The development set was used to determine the optimal SVM parameters in each case, which are then used for the classification applied to the test set. In separate experiments, TC2015 was used as the test set, keeping the same subsets of TC2014 as training sets, to make the experiments comparable and to assess the usefulness of a year-old tweets to classify new tweets. Keeping the same subsets of TC2014 as training sets enables us to compare how well the classifier does with new tweets by using exactly the same model for training.

We created eight different SVM classifiers, each of which used one of the following eight features available from a tweet as retrieved from a stream of the Twitter API:

1) **User location (uloc):** This is the location that the user specifies in their profile. While this feature might seem a priori useful, it is somewhat limited as this is a free text field that users can leave empty, input a location name that is ambiguous or has typos, or any other string that does not match with any specific locations (e.g., “at home”). Looking at the users’ self-reported locations, Hecht et al. [18] found that 66% report information that can be translated, accurately or inaccurately, to a geographic location, with the other 34% being either empty or not geolocalisable.

2) **User language (ulang):** This is the language in which the user uses Twitter, i.e., the users self-declared user interface language. The interface language that a user is using might be indicative of their country of origin; however, they might have set up the interface in a language different from theirs, such as English, for instance because it was the default language when they signed up or because the language of their choice is not available. Another issue might be that a language is often not unique to a single country, and hence it can be ambiguous at times, as is indeed the case with widely used languages such as English and Spanish.

3) **Timezone (tz):** This indicates the time zone that the user has specified in their settings, e.g., “Pacific Time (US & Canada)”. When the user has specified an accurate time zone in their settings, it can be indicative of their country of origin; however, some users may have the default time zone in their settings, or they may use a time zone that is equivalent to their but referring to a different location (e.g., “Europe/London” for a user in Portugal). Also, Twitter’s predefined list of time zones does not include all the countries.

4) **Tweet language (tlang):** This is automatically detected by Twitter and refers to the language in which the tweet is believed to be written. It has been found to be accurate for major languages, but it leaves much to be desired for less widely used languages, often being inaccurate owing to the difficulty of determining the language of such a short text. Twitter’s language identifier has also been found to struggle with multilingual tweets, where parts of a tweet are written in different languages [53].

5) **Offset (offset):** This is the offset, with respect to UTC/GMT, that the user has specified in their settings. This feature is similar to the time zone, albeit more limited as it only refers to the local time, which is often shared by a number of countries.

6) **Name of the user (name):** This is the name that the user specifies in their settings, which can be their real name, or an alternative name they choose to use.

---

6. http://www.geonames.org/
The name of a user can reveal, on some occasions, their country of origin.

7) **User description (description):** This is a free text where a user can describe themselves, their interests, etc. It can be empty if the user opts not to fill it, or it can be a string that ranges from descriptive content that might say something about the user’s origin to text that is useless for the purposes of determining their country of origin.

8) **Tweet content (content):** The text that forms the actual content of the tweet. The use of content has a number of caveats. One is that content might change over time, and therefore new tweets might discuss new topics that the classifiers have not seen before. Another caveat is that the content of the tweet might not be relevant to their location, or the tweet might not refer to any locations; in a previous study, Rakesh et al. [41] found that the content of only 289 out of 10,000 tweets was location-specific.

Figure 2 shows an example of a tweet and the eight features listed above. The features were treated in two different ways: the user location, name of the user, description and tweet content were represented using a bag of words approach, where each token represented a feature in the vector space model. The rest of the features, namely the user language, time zone, tweet language and offset, were represented by a single categorical value in the vector space model, given the limited number of values that the features can take. We used these eight features separately, as well as in different combinations with one another, in our experiments testing the ability to infer the country of origin of tweets. As we noted above, we used SVM committees to test different combinations of these features.

4.3 Evaluation

We report three different performance values for each of the experiments: micro-accuracy, macro-accuracy and mean squared error (MSE). The accuracy values are computed as the result of dividing all the correctly classified instances by all the instances in the test set. The micro-accuracy is computed for the test set as a whole. The macro-accuracy penalises accuracy in the whole dataset, the macro-accuracy computed for the test set as a whole. For the macro-accuracy, the result of dividing all the correctly classified instances by all the instances in the test set. The micro-accuracy is the actual number of correctly classified instances.

Having the latitude and longitude values of the centroids of all these countries, we used the Haversine formula [43], which accounts for the spheric shape when computing the distance between two points and is often used as an acceptable approximation to compute distances on the Earth. The Haversine distance between two points of a sphere each defined by its longitude and latitude is computed as shown in Equation 4.

$$d = 2r \arcsin \left( \sqrt{\sin^2 \left( \frac{\varphi_2 - \varphi_1}{2} \right) + \cos(\varphi_1) \cos(\varphi_2) \sin^2 \left( \frac{\lambda_2 - \lambda_1}{2} \right)} \right)$$

where $\varphi_1$ and $\varphi_2$ are the latitudes of point 1 and point 2, $\lambda_1$ and $\lambda_2$ are the longitudes of point 1 and point 2, and $r$ is the radius of the Earth, which is estimated to be 6,371 km.

5 Classification Results

In this section, we present results for different location classification experiments. First, we look at the performance of classifiers that use a single feature. Then, we present the results for classifiers combining multiple features. To conclude, we examine the results in more depth by looking at the performance by country, as well as error analysis.

5.1 Single Feature

Table 2 shows the results for the classification on the TC2014 dataset with two different approaches using GeoNames, one based on population (the most populous city is chosen when there are different options for a name) and one based on relevance (the city name that most resembles the input string). In this dataset, 65.82% of the tweets have a non-empty string in the location field; for the rest of tweets, the GeoNames-based approach will not be able to determine the country. For a fair evaluation with respect to our approaches, we pick the most popular country in the dataset as the output of the approach based on GeoNames. The table shows values of micro- and macro-accuracy.

| Feature       | Microacc. | Macroacc. | MSE      |
|---------------|-----------|-----------|----------|
| population    | 0.505     | 0.317     | 1505.661 |
| relevance     | 0.504     | 0.342     | 1505.586 |

Table 3 shows the results of SVMs, each making use of only one of the eight features under study. This table includes performance values when we applied the classifier on both datasets, TC2014 and TC2015. The additional column, “Diff.”, shows the relative difference in performance for each of these datasets, i.e., measuring the extent to which a model learned from the TC2014 dataset can still be applied
It is absolutely gorgeous outside. We will be delivering ice cream all day if you feel the need to not step out.

Fig. 2. Example of a tweet and the 8 features that we used to infer the country of origin.

| Feature       | Microaccuracy | Macroaccuracy | MSE       |
|---------------|---------------|---------------|-----------|
| content       | 0.418         | 0.336         | -19.6%    |
| description   | 0.319         | 0.310         | -2.8%     |
| name          | 0.398         | 0.406         | +2.0%     |
| offset        | 0.369         | 0.321         | -13.0%    |
| tlang         | 0.553         | 0.523         | -5.4%     |
| tz            | 0.438         | 0.431         | -1.6%     |
| ulang         | 0.437         | 0.428         | -2.1%     |
| uloc          | 0.453         | 0.511         | +12.8%    |

TABLE 3

Classification results with SVMs on a single feature for all the countries in TC2014 and TC2015.

to the TC2015 test set. Note that while higher values are desired for microaccuracy and macroaccuracy, lower values are optimal for MSE.

Two approaches stand out here. On the one hand, tweet language (tlang) is the best performing feature in terms of microaccuracy and this can also be observed in the lowest error rates in terms of MSE. We believe that this is due to the proximity of many countries that speak the same language (e.g., Germany and Austria, or Argentina and Chile), in which cases the classifier that only relies on tweet language can confuse them while the distance is relatively small. On the other hand, the user’s self-reported location (uloc) performs by far the best in terms of macroaccuracy. This is due to the fact that the classifier that only uses the user’s profile location will be able to guess correctly a few cases for each country where users specify a correctly spelled, unambiguous location, but will fail to classify correctly the rest of the cases; hence the higher macroaccuracy is sensible according to these expectations, and also in line with the performance results by using GeoNames. While most of these classifiers outperform the GeoNames baseline in terms of microaccuracy, the best feature (uloc) is still behind the baseline in terms of macroaccuracy, which suggests that alternative approaches are needed for a better balanced classification performance.

Figure 3 shows a heat map with accuracy values of each of the features broken down by country. We observe the best distributed accuracy across countries is with the use of user location as a feature. However, other features are doing significantly better classifying tweets that belong to some of the major countries such as the USA (better classified by tweet language or time zone, among others), Russia (better classified by tweet language) or Brazil (better classified by tweet language or name, among others). This emphasises the necessity to further dig into the differences among each country’s characteristics, which our work aims to explore.

As we noted above, a remarkable characteristic of our datasets (and the reality of Twitter itself) is the high imbal-
Fig. 3. Accuracy by country for each of the eight features used alone in the SVM classifier.
ance in the distribution of tweets across countries, where a few countries account for a large majority of the tweets and many countries in the tail account for very few tweets. The fact that the classifier has to determine which of the 217 countries a tweet belongs to substantially complicates the task. To quantify this, and to explore the ability to boost performance on the countries with highest presence, we also performed classification experiments on the top 25 countries. These top 25 countries account for as many as 90.22% of the tweets; consequently, being able to boost performance on these 25 countries, while assuming that the system will miss the rest, can make it a more achievable task where the overall performance gets improved.

To perform the classification on the top countries, we removed the tweets pertaining to countries that do not belong to the top 25 list from the training set, so that the system will only train its model from tweets pertaining to those countries. Including all the other tweets from the remaining countries would add a noisy category to the training set, given the diversity of that new category. However, we would not do the same for the test set, as one will still observe tweets that do not belong to those 25 countries in the application of the classifier. For the purposes of the experimentation, we assign the rest of the tweets in the test set a different, 26th label, meaning that they belong to other countries. Our experiments on the top 25 countries will then have a training set with 25 categories to learn from, and test sets with 26 categories, where the classifier will never predict the 26th category.

Table 4 shows the results for the experiments on the top 25 countries. The overall tendency is very similar to that of the SVM classifiers applied to all the countries in the world, where the tweet language (tlang) leads to best performance in terms of microaccuracy and MSE, and the user location (uloc) performs by far best in terms of macroaccuracy. Surprisingly, microaverage results are very similar, 0.553 and 0.550 for all countries and top 25 respectively in TC2014, and 0.523 and 0.521 in TC2015. However, the striking difference can be observed especially in the macroaccuracy values, indicating that the macroaccuracy results for all countries are heavily affected by the least represented countries. While the classifier using user location achieved around 0.28 in terms of macroaccuracy for all the countries, this value increases dramatically to 0.421 and 0.467 for TC2014 and TC2015 when we apply the same classifier to the top 25 countries. Classification on a reduced subset of countries can substantially boost performance, even assuming that part of the dataset will be misclassified. In fact, classification on this optimised setting can also outperform the baseline using GeoNames. Not only does the top performing feature, user location, improve its performance. Other features that performed poorly before, such as tweet language, time zone or user language, perform significantly better, nearly achieving the same performance level as the GeoNames baseline. This further motivates our subsequent goal of studying combinations of features to further boost the performance of the classifier applied to the top 25 countries.

### 5.2 Feature Combinations

Having seen that different features give rise to gains in different ways, testing the performance of combinations of multiple features seemed like a wise option. As we described in subsection 4.1, we performed these combinations of features by using SVM committees that aggregate the output of different SVM classifiers each using a different feature. We tested all 255 possible combinations using the eight features under study. We only report the best performing combinations here in the interest of space and clarity.

Table 5 shows the best combination in each case, both for the TC2014 and TC2015 datasets, as well as for the classifiers that consider all the countries in the datasets and only the top 25 countries. The three best approaches in each case correspond to the best in terms of microaccuracy, the best in terms of macroaccuracy, and the best in terms of MSE, each of which have the corresponding column highlighted in bold. We observe that the selection of an appropriate combination of features can actually lead to a substantial increase in terms of microaccuracy, tweet language and user language being common features for the classifiers performing best in terms of microaccuracy in all the settings. Microaccuracy can be boosted with a combination of features from the previous best of 0.550 to microaccuracy values close to 0.6. However, when one wants to boost macroaccuracy, the classifier that solely relies on the user’s location yields the best performance, and a combination with any other feature harms its performance. Surprisingly, while user location is not always defined by the user or it can be a string that does not reveal the user’s actual location, it still achieves the best balanced performance across countries. Finally, if we aim to optimise MSE, the combination of multiple features will lead to the best performance, among which tweet language, time zone and user location are the common features that are helpful in all settings.

Interestingly, the combination of features has led to a significant improvement in terms of microaccuracy, which is still not better balanced across countries. This leads us to think about the differences that countries may have among them. Will different sets of features be useful for an accurate classification for each country? Are we perhaps doing very well for some countries with certain combinations, but that combination, in turn, is bad for other countries? To further dig into this, we will now take a closer look at the performance broken down by country.

### 5.3 Breakdown of Countries

Given the remarkable differences among countries we observed in Figure 3 when exploring how different features are useful for different countries, and that combinations of features struggle to outperform the sole use of user location as a feature for the classification, we take a closer look at the performance of different SVM committees for each of the top 25 countries. As we are now looking at each country separately, we use precision, recall and F1 scores as more appropriate evaluation measures that better capture the extent to which a country’s instances are being correctly categorised. We look at the best combination of features for each country in terms of F1 score, and analyse the set of features that lead to the best performance in each case. We show the results of this analysis in Table 6.

The results show that very different approaches lead to optimal results for each country, revealing the different fea-
**TABLE 5**

Performance results for combinations of features using SVMs.

| Country      | Best SVM combination | Performance | Best SVM combination | Performance |
|--------------|----------------------|-------------|----------------------|-------------|
|              | content              | name        | offset               | lng         | ulang       | tlang       | uloc       | P   | R   | F1   |
| Chile        |                       | 0.844       | 0.579               | 0.687       |             |             |             | 0.824 | 0.608 | 0.700 | +1.9% |
| Spain        |                       | 0.725       | 0.445               | 0.552       |             |             |             | 0.634 | 0.495 | 0.556 | +0.7% |
| Turkey       |                       | 0.859       | 0.984               | 0.971       |             |             |             | 0.984 | 0.984 | 0.974 | +0.3% |
| Brazil       |                       | 0.892       | 0.964               | 0.927       |             |             |             | 0.865 | 0.954 | 0.908 | -2.0% |
| India        |                       | 0.816       | 0.481               | 0.605       |             |             |             | 0.820 | 0.526 | 0.641 | +6.0% |
| Saudi Arabia |                       | 0.687       | 0.685               | 0.866       |             |             |             | 0.498 | 0.513 | 0.505 | -26.4%|
| Mexico       |                       | 0.860       | 0.431               | 0.574       |             |             |             | 0.855 | 0.466 | 0.603 | +5.1% |
| Netherlands  |                       | 0.569       | 0.610               | 0.680       |             |             |             | 0.609 | 0.484 | 0.539 | -20.7%|
| Thailand     |                       | 0.974       | 0.773               | 0.862       |             |             |             | 0.972 | 0.436 | 0.621 | -28.0%|
| UK           |                       | 0.747       | 0.543               | 0.629       |             |             |             | 0.614 | 0.606 | 0.610 | -3.0% |
| Colombia     |                       | 0.879       | 0.252               | 0.391       |             |             |             | 0.802 | 0.299 | 0.436 | +11.5%|
| South Africa |                       | 0.742       | 0.348               | 0.630       |             |             |             | 0.797 | 0.563 | 0.660 | +8.8% |
| Australia    |                       | 0.593       | 0.530               | 0.637       |             |             |             | 0.527 | 0.543 | 0.643 | +10.0%|
| Indonesia    |                       | 0.729       | 0.919               | 0.852       |             |             |             | 0.734 | 0.910 | 0.813 | -4.6% |
| USA          |                       | 0.957       | 0.915               | 0.692       |             |             |             | 0.729 | 0.606 | 0.662 | -4.3% |
| France       |                       | 0.812       | 0.867               | 0.839       |             |             |             | 0.677 | 0.717 | 0.697 | -16.9%|
| Japan        |                       | 0.956       | 0.976               | 0.966       |             |             |             | 0.946 | 0.961 | 0.959 | -1.3% |
| Canada       |                       | 0.956       | 0.502               | 0.603       |             |             |             | 0.872 | 0.553 | 0.638 | +5.8% |
| Italy        |                       | 0.908       | 0.818               | 0.861       |             |             |             | 0.864 | 0.745 | 0.800 | -7.1% |
| Argentina    |                       | 0.851       | 0.290               | 0.433       |             |             |             | 0.811 | 0.324 | 0.463 | +6.9% |
| Venezuela    |                       | 0.814       | 0.461               | 0.588       |             |             |             | 0.825 | 0.481 | 0.608 | +3.4% |
| Russia       |                       | 0.530       | 0.971               | 0.820       |             |             |             | 0.664 | 0.932 | 0.776 | -5.4% |
| Germany      |                       | 0.639       | 0.531               | 0.580       |             |             |             | 0.620 | 0.489 | 0.546 | -5.9% |
| Philippines  |                       | 0.789       | 0.480               | 0.597       |             |             |             | 0.855 | 0.467 | 0.604 | +1.2% |
| Malaysia     |                       | 0.502       | 0.637               | 0.562       |             |             |             | 0.548 | 0.608 | 0.576 | +2.5% |

Classification results broken down by country for SVM committees on the top 25 countries. The color code represents how the best sets of features for TC2015 compare to those for TC2014 (blue: countries where the same set of features works best for TC2014 and TC2015; green: countries where a reduced set of features from TC2014 works best for TC2015; red: countries where new features, not used in the best approach for TC2014, works best for TC2015).

tures that characterise each country and differentiate them from the rest. The simple use of a single feature is actually best in some cases, especially the user location. This is true for countries like the United Kingdom, which may share many of the other features with the United States, but a large number of users seem to specify a valid location in their profile, and leads to this being the best feature. The use of the user location is also best for Chile, Spain, India, Mexico, South Africa, Canada and Philippines. Most of these cases are actually consistent in TC2014 and TC2015, suggesting that when user location works well alone, it guarantees good performance in a long term too; the exception is Canada, where the inclusion of the name improves performance in TC2015. However, other countries benefit from having many features combined. For the United States and Indonesia, the best results are achieved by using as many as 7 features. Other countries like Turkey, Netherlands, Japan, Italy and Venezuela use 4 features. Yet, the number of features that need to be used decreases significantly from TC2014 and TC2015, showing that the usefulness of all features does not persist at the same level. This is further corroborated when we look at the total number of features used in TC2014 and TC2015. While overall the 25 countries use 64 features in the former, this number drops to 49 in the latter. In fact, the number of countries in which each of the features is used drops in most cases, except for description and user location, where it remains constant. Thankfully, in 16 out of the 25 countries, the same combination of features performs best, while 5 countries use a reduced subset of features from TC2014 in TC2015, and only 4 countries change the set of features including new ones. Hence, one has to be careful when choosing the features to be included in the classifier as the trained model gets old, but will rarely have to consider adding new features that had not been
considered before.

One aspect that stands out when looking at the features is that content, one of the features that is most commonly used in previous works for geolocation purposes, is useful in very few cases in TC2014, while it is never useful in TC2015. On the one hand, this suggests that there are features that generally lead to more accurate classification than content and, on the other hand, this shows that the validity of a model trained on content is time dependent and is no longer as useful when people talk about different things later. Instead of content, the features that are most useful after a while include user location (in 15 countries), user language (9), name (9) and tweet language (6).

When we look at the varying accuracy values by country, we observe very high performance values for large countries that have either an own language or a language that is predominantly spoken in that country: Turkey, Brazil, Thailand, Indonesia, France, Japan, Italy and Russia. Interestingly, the best approach for all of these countries include either or both of tweet language or user language. When it comes to user language, this means that users in these countries have a strong inclination towards setting the user interface in their own language instead of the default language. In the case of tweet language, this mainly reflects a combination of two things, one being that users in this country tend to tweet mostly in their own language, while the other is that Twitter’s automated language identification system is very accurate in these cases.

However, at the low performance end we have countries for which we have not managed to achieve accuracy values that are as high as for others. This is especially the case for English and Spanish speaking countries, where the classifier has struggled with there being several countries in each of these cases, which share similar content, in the same language, and have names in the same language.

Finally, looking at the performance difference of countries in TC2014 and that in TC2015, there is no big gap in most of the cases and the differences are mostly within ±6%. However, there are a few cases where the performance drops drastically when we apply the classifier on the new dataset. This is the case of Saudi Arabia, Netherlands, Thailand and France, whose performance in TC2015 drops between 16% and 26% from that in TC2014. The highest improvement occurs for Colombia, with 11% increase in performance in TC2015.

5.4 Error Analysis
To shed some light on the reasons why some countries are not classified as accurately, we look at the errors that the classifiers are making. Overall, if we put together all correct classifications by any of the classifiers, we would be able to get a microaccuracy of up to 97.3% as an upper bound estimation for the tweets that belong to one of the top 25 countries. This puts high expectations in that nearly all users can be accurately classified in some way by using the right classifier. However, many countries share similar (or common) characteristics, which often leads to mistakes between those countries. To better understand this, we look at the confusion matrix for the top 25 countries.

The confusion matrix in Table 7 shows the aggregated deviations between predictions and ground truth for all the 255 classifiers applied to the top 25 countries. The values highlighted in gray refer to correct guesses (diagonal) and deviations towards the 26th category for the rest of the countries (last column). In red, we highlight the most prominent deviations exceeding 10% of a country’s instances, and in orange those exceeding 5% of the instances.

On the positive side, some of the countries do not deviate 5% or more to any other country (i.e., no red or orange cells), including Australia, Germany, Japan, Netherlands, Russia and Turkey. However, a striking observation we make from this confusion matrix is the large amount of deviations between Spanish speaking countries, which include Argentina, Chile, Colombia, Spain, Mexico and Venezuela. In most of these cases, in fact, the high number of confusions occurs in both directions for each pair of countries. This is indeed an additional difficulty that one might have expected, given that all of them share cultural and linguistic commonalities, especially for using the same language and hence overlapping content. Moreover, the Latin American countries often share the time zone and, while the time zone is different for Spain, many of the cities in the Latin American countries are named after Spanish cities (e.g., Córdoba in Argentina, León in Mexico, Valencia in Venezuela, Cartagena in Colombia or Santiago in Chile, all of which are also Spanish cities), which makes the distinction from Spain more challenging if only the user location is used. Additionally, many countries deviate towards the United States, most probably for being the predominant country on Twitter and accounting for about 20% of the tweets in the datasets. The remaining cases with large numbers of deviations include cases in which the tweet is actually posted from South Africa, but the classifier thinks it is from Indonesia or Philippines. While one might think that these three countries nothing in common a priori, further digging into the data enables us to observe remarkable commonalities between them. In fact, a majority of the users in these countries share the following characteristics: (1) they have the user interface set up in English, (2) they tweet mostly in English, and (3) most of them have the time zone and UTC offset set to “None”. In cases like this, users with the default settings are very difficult to categorise accurately.

6 Discussion
Our experiments and analysis on over 5 million geolocated tweets from unique users reveal insight into the development of an automated classifier for country-level geolocation of tweets. Our experiments only make use of features inherent in the tweets as retrieved through Twitter’s streaming API, compatible with a scenario of classifying tweets by country in real-time. This can be invaluable when curation of the tweet stream is needed for applications such as country-specific trending topic detection, or for more specific applications where only tweets coming from a specific country are sought. The identification of the country of origin will also help mitigate problems caused by the limited availability of demographic details for Twitter users.

We found that one of the straightforward and most commonly used approaches, which is the use of gazetteers such as GeoNames to match the user’s self-reported location with
Aggregated confusion matrix for all SVMs on the top 25 countries. (ar: Argentina, au: Australia, br: Brazil, ca: Canada, cl: Chile, co: Colombia, de: Germany, es: Spain, fr: France, gb: United Kingdom, id: Indonesia, in: India, it: Italy, jp: Japan, mx: Mexico, my: Malaysia, nl: The Netherlands, ph: Philippines, ru: Russia, sa: Saudi Arabia, th: Thailand, tr: Turkey, us: United States, ve: Venezuela, za: South Africa)
a place in the world, performs well in terms of macroaccuracy, correctly identifying tweets for many countries, but fails in terms of microaccuracy, doing only slightly well but without high accuracy for most of the countries. The use of a SVM classifier that solely makes use of the language of the tweet substantially outperforms the GeoNames baseline in terms of microaccuracy. However, the use of a single feature in a classifier cannot improve the performance of GeoNames in terms of macroaccuracy. The main challenge of the classifier is that it has to deal with as many as 217 countries, making the task especially difficult.

To overcome this, we have tested our classifier on a reduced subset of the top 25 countries, which still account for more than 90% of the whole Twitter stream. In this case, we found that the classifier can substantially outperform the GeoNames baseline. The sole use of tweet language performs significantly better in terms of microaccuracy, while we found it more appropriate to use the user’s self-reported location as a feature when the objective is to achieve high macroaccuracy.

Further testing with combinations of multiple features, we found that this performance can still be improved, although one needs to be careful when picking the features to be used, depending on whether the main objective is high microaccuracy or macroaccuracy. What is interesting is that the a classifier trained on data from the same time frame as the test set can be effectively applied to new tweets, which we verified on tweets posted a year later. The combination of features that works well for the test set in the same time frame can be applied to the new tweets in most cases, achieving similar performance values. However, we observed that it is better to disregard tweet content when one wants to classify new tweets, as the validity of the model trained in content drops. While previous work has found that tweets can be accurately geolocated based on content for tweets within the same time frame, our study shows that this is no longer the case for new tweets.

The scenario is quite different when one wants to identify tweets from a specific country, given that different sets of features lead to more accurate classifications for different countries, which do not necessarily match with the overall best approach. By picking the right combination of features one can achieve classification performances for a country higher than 0.9 in terms of F1 score when it is a large country and has unique characteristics such as having a language that is not spoken in many other countries or having a unique time zone. However, these performance values tend to drop when one aims to identify tweets for a country that has common characteristics with other countries; this is especially true for English and Spanish speaking countries, among which many are large countries that speak the same language, share similar contents, and have the same time zone (e.g., Chile and Argentina).

In summary, the results are sufficiently promising so as to conclude that an appropriate selection of the features inherent to a tweet can lead to accurate, real-time classification of the most populous countries in terms of Twitter users. This can indeed be applied not only to tweets collected within the same time frame as the tweets used to train the model, but it can also be applied to tweets collected later in time when the topics that users talk about are expected to be completely different.

7 Conclusion

To the best of our knowledge, this is the first study performing a comprehensive analysis of the usefulness of tweet-inherent features to automatically infer the country of origin of tweets in a real-time scenario. Most of the previous work had focused on classifying tweets coming from a single country and hence assuming that content from that country was already identified. Where previous work had considered tweets from all over the world, however, the set of features employed for the classification included features that are not readily available within a tweet, such as a user’s social network, which is not applicable in a scenario where tweets need to be classified in real time as they are collected from the streaming API. Our study presents, as far as we know, the first work that deals with a worldwide stream of tweets written in any language and that looks at how different tweet-inherent features can help with country-level geographical inference of tweets. Beyond that, our study uses two datasets collected a year apart from each other, to test the ability to classify new tweets with a classifier trained on older tweets. Our experiments and analysis reveal insight that can be effectively used to build an application that classifies tweets by country in real time, either when the goal is to organise content by country or when one wants to filter out all the content that does not belong to a specific country.

In the future we plan to test more sophisticated approaches for content analysis to see if there is an alternative way to make content useful in a long term. Possible alternatives to better exploit the content of a tweet, moving away from the use of a bag-of-words approach, can be the detection of categories topics in content (e.g. do some countries talk more about football than others?), as well as a semantic treatment of the content.

Acknowledgments

This work has been supported by the PHEME FP7 project (grant No. 611233), the Warwick University Higher Education Impact Fund, an ESRC Impact Acceleration Award, EPSRC Impact Acceleration Account (grant no. EP/K503940/1) and EPSRC grant EP/L016400/1.

References

[1] O. Ajao, J. Hong, and W. Liu. A survey of location inference techniques on twitter. Journal of Information Science, 1:1–10, 2015.
[2] H. Bo, P. Cook, and T. Baldwin. Geolocation prediction in social media data by finding location indicative words. Proceedings of COLING 2012, pages 1045–1062, 2012.
[3] J. Bollen, H. Mao, and A. Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. Proceedings of ICWSM, 11:450–453, 2011.
[4] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In COLT ’92: Proceedings of the fifth annual workshop on Computational learning theory, pages 144–152, New York, NY, USA, 1992. ACM.
[5] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella. Discriminating gender on twitter. In Proceedings of EMNLP, pages 1301–1309, 2011.
[6] Y. Chen, J. Zhao, X. Hu, X. Zhang, Z. Li, and T.-S. Chua. From interest to function: Location estimation in social media. In Proceedings of the AAAI Conference on Artificial Intelligence, 2013.
