# Classification of Worldwide Tweets in Real-time based on Location

[1] *PYLA TULASI,* [2] *Prof. B.PRAJNA*
[1,2] *Department of Computer Science and System Engineering,*
*Andhra University College of Engineering (A), Visakhapatnam, AP, India*

***Abstract****: Social media are increasingly being used in the scientific community as a key source of data to help understand diverse natural and social phenomena, and this has prompted the development of a wide range of computational data mining tools that can extract knowledge from social media for both post-hoc and real time analysis. Twitter has become a leading data source for such studies. 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 location-specific trending topics, emerging events and breaking news. In contrast to much previous work that has focused on location classification of tweets restricted to a specific country, here we undertake the task in a broader context by classifying global tweets at the country level, which is so far unexplored in a real-time scenario. We analyze the extent to which a tweet's country of origin can be determined by making use of eight tweet-inherent features for classification. Furthermore, we use two datasets, collected a year apart from each other, to analyze the extent to which a model trained from historical tweets can still be leveraged for classification of new tweets.*

***Index Terms*** *– Twitter, Geo-Location, Real-Time, Classification*

## 1 INTRODUCTION

The increase of interest in using social media as a source for research has motivated tackling the challenge of automatically geo-locating tweets, given the lack of explicit location information in the majority of tweets. In contrast to much previous work that has focused on location classification of tweets restricted to a specific country, here we undertake the task in a broader context by classifying global tweets at the country level, which is so far unexplored in a real-time scenario. We analyze the extent to which a tweet's country of origin can be determined by making use of eight tweet-inherent features for classification [1]. Furthermore, we use two datasets, collected a year apart from each other, to analyze the extent to which a model trained from historical tweets can still be leveraged for classification of new tweets.

Social media are increasingly being used in the scientific community as a key source of data to help understand diverse natural and social phenomena, and this has prompted the development of a wide range of computational data mining tools that can extract knowledge from social media for both post-hoc and real time analysis. Thanks to the availability of a public API that enables the cost-free collection

of a significant amount of data, Twitter has become a leading data source for such studies [2].

Having Twitter as a new kind of data source, researchers have looked into the development of tools for real-time trend analytics or early detection of newsworthy events as well as into analytical approaches for understanding the sentiment expressed by users towards a target or public opinion on a specific topic. However, Twitter data lacks reliable demographic details that would enable a representative sample of users to be collected and/or a focus on a specific user subgroup or other specific applications such as helping establish the trustworthiness of information posted.

Automated inference of social media demographics would be useful, among others, to broaden demographically aware social media analyses that are conducted through surveys[3]. One of the missing demographic details is a user's country of origin, which we study here. 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. Most of the previous research in inferring tweet geolocation has classified tweets by location within a limited geographical area or country; these cannot be applied directly to an unfiltered stream where tweets from any location or country will be observed [4,5,6] . The few cases that have dealt with a global collection of tweets have used an extensive set of features that cannot realistically be extracted in a real-time, streaming context (e.g., user tweeting history or social networks) and have been limited to a selected set of global cities as well as to English tweets.

This means they use ground truth labels to pre-filter tweets originating from other regions and/or written in languages other than English[8]. The classifier built on this pre-filtered dataset may not be applicable to a Twitter stream where every tweet needs to be geo-located. 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 location-specific trending topics[14], emerging events and breaking news.

Specific applications of a real-time, country-level tweet geolocation system include country-specific trending topic detection or tracking sentiment towards a topic broken down by country [9,12]. 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.

## 2 EXISTING SYSTEM

Most of the previous studies on automated geo-location of tweets have assumed that the tweet

stream includes only tweets from a specific country . 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. Another widely used technique relies on the social network that a user is connected to, in order to infer a user's location from that of their followers and followees . 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.

## 3 DRAWBACKS

- Location is identified only if GPS is on but it is not necessary to turn on GPS while using Facebook.
- User's Location is only identified by the user's self reported location.

## 4 PROPOSED SYSTEM

In the proposed work the increasing interest in inferring the geographical location of either messages or twitter users. The automated inference of tweet location has been studied for different purposes, ranging from data journalism, to public health. The summary of previous work reported in the scientific literature, outlining the features that each study used to classify messages 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 messages, aggregated multiple messages and/or user history were used to train the classifier.

## 5 METHODOLOGY

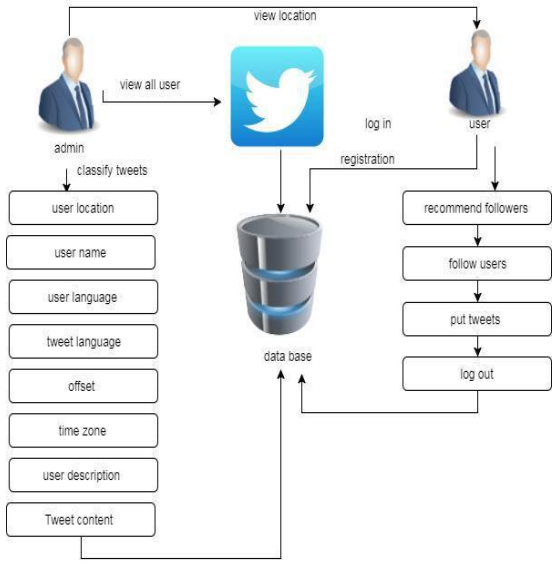


**Figure1**: System architecture

In this study, a country-level location classification task is defined as one in which, given a single tweet as input, a classifier has to determine the country of origin of the tweet. The content and metadata provided in a single tweet are considered, 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 network derivable from a user's followers and followers, which would not be feasible in our real-time scenario.

**Classification Techniques**

The classification is carried out with a range of classifiers of different types: Support Vector Machines (SVM), Naive Bayes, Decision Trees, Random Forests and a Maximum Entropy classifier. The classifier uses one of the eight features available from a tweet as retrieved from a stream of the Twitter API:

a) User location (uloc): This is the location the user specifies in their profile.

b) User language (ulang): This is the user's self-declared user interface language.

c) Timezone (tz): This indicates the time zone that the user has specified in their settings, e.g., "Pacific Time (US & Canada)".

d) Tweet language (tlang): The language in which a tweet is believed to be written is automatically detected by Twitter.

e) Offset (offset): This is the offset, with respect to UTC/GMT, that the user has specified in their settings.

f) User name (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.

g) User description (description): This is a free text where a user can describe themselves, their interests, etc.

h) Tweet content (content): The text that forms the actual content of the tweet.

## 5 EVALUATION

Three various performance values for each of the experiments: micro-accuracy, macro-accuracy and mean squared error (MSE) are used. 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. For macro-accuracy, we compute the accuracy for each specific country in the test set, which are then averaged to compute the overall macro-accuracy. While the micro-accuracy measures the actual accuracy in the whole dataset, the macro-accuracy penalises the classifier that performs well only for the majority classes and rewards, instead, classifiers that perform well across multiple categories. The MSE is the average of the squared distance in kilometers between the predicted country and the actual, ground truth country.

## 7 RESULTS

The experiments and analysis on over geo located tweets from unique users reveal insight into country-level geo location of tweets in real time. Country- specific trending topic detection can be found such that which topics have the

highest number of tweets in which location. more specific applications where only tweets coming from a specific country are sought, e.g., sentiment analysis[15] where the tweets are analyzed and are classified as positive, negative and neutral. The identification of the country of origin will also help mitigate problems caused by the limited availability of demographic details for Twitter users.

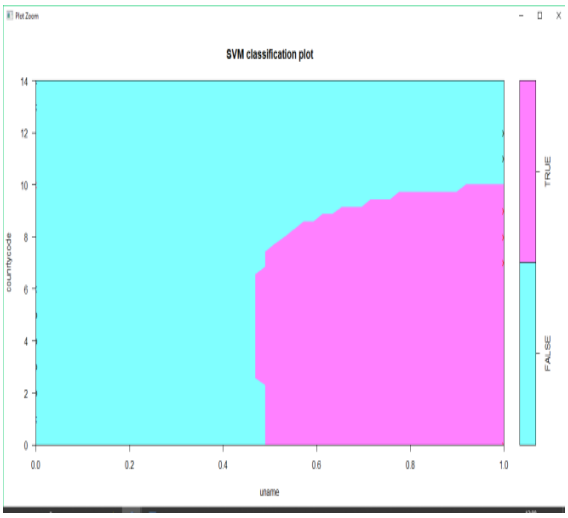


**Figure2:**SVM classification plot

**Performance of Various Classification Algorithms:**

| Algorithm | Accuracy |
|---|---|
| SVM Classification | 0.07142857 |
| Naïve Bayes | 0.354354 |
| Decision Tree Prediction | 0.1428571 |
| Random Forest classification | 0.09285714 |

## 8 CONCLUSION

To the simplest of my data, this is the primary study acting a comprehensive analysis of the quality of tweet inherent options to infer the Country of origin of tweets in real time from a worldwide stream of tweets written in any language. Most previous work focused on classifying tweets coming from one country and therefore assumed that tweets from that country were already known. wherever previous work had thought of tweets from everywhere the globe, the set of options used for the classification enclosed options, similar to a user's social network, that don't seem to be without delay accessible within a tweet then isn't possible in a very situation wherever tweets have to be compelled to be classified in real time as they're collected from the streaming API.. Finally, our study uses two datasets collected a year except one another, to check the flexibility to classify new tweets with a classifier trained on older tweets. Our experiments and analysis reveal insights which will be used effectively to create an application that classifies tweets by country in real time, either once the goal is to arrange content by country or one desires to spot all the content from a particular country.

## 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] E. Amig´ o, J. C. De Albornoz, I. Chugur, A. Corujo, J. Gonzalo, T. Mart´ın, E. Meij, M. De Rijke, and D. Spina. Overview of replab 2013: Evaluating online reputation monitoring systems. In Proceedings of CLEF, pages 333–352. Springer, 2013.

[3] F. Atefeh and W. Khreich. A survey of techniques for event detection in twitter. Computational Intelligence, 31(1):132–164, 2015.

[4] H. Bo, P. Cook, and T. Baldwin. Geolocation prediction in social media data by finding location indicative words. In Proceedings of COLING, pages 1045–1062, 2012.

[5] J. Bollen, H. Mao, and A. Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In Proceedings of ICWSM, pages 450–453, 2011.

[6] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella. Discriminating gender on twitter. In Proceedings of EMNLP, pages 1301–1309, 2011.

[7] H.-w. Chang, D. Lee, M. Eltaher, and J. Lee. @ phillies tweeting from philly? predicting twitter user locations with spatial word usage. In Proceedings of ASONAM, pages 111–118, 2012.

[8] 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 AAAI, pages 180–186, 2013.

[9] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In Proceedings of CIKM, pages 759–768, 2010.

[10] R. Compton, D. Jurgens, and D. Allen. Geotagging one hundred million twitter accounts with total variation minimization. In IEEE Big Data, pages 393–401, 2014.

[11] M. Conover, J. Ratkiewicz, M. R. Francisco, B. Gonc¸alves, F. Menczer, and A. Flammini. Political polarization on twitter. In Proceedings of ICWSM, pages 89–96, 2011.

[12] M. D. Conover, B. Gonc¸alves, J. Ratkiewicz, A. Flammini, and F. Menczer. Predicting the political alignment of twitter users. In IEEE PASSAT/SocialCom, pages 192–199, 2011.

[13] D. Doran, S. Gokhale, and A. Dagnino. Accurate local estimation of geo-coordinates for social media posts. arXiv preprint arXiv:1410.4616, 2014.

[14] B Prajna, N Sneha Application for retrieving details of users - Topic based approach, IJCSETpages 509-513,2015

[15] B. Prajna.K.Indu. Sentiment Analysis for Twitter Real Time Tweets, 2015.