International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

Automatic geo locating for tweets

[1]**K.Praveen Kumar**, Praveen.koram@gmail.com,

[2]**T.L.Shreya,** Shreyastrovey48@gmail.com

[3]**S.Naresh Reddy,** Nareshsama143@gmail.com

[4]**V.Vinod Kumar,** Vinodvanam568@gmail.com

[1, 2, 3,4]Dept.of Computer Science and Engineering, Vignan Institute of Technology and Science, Deshmukhi, Hyderabad. 508284

***Abstract**:*

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. 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. With classification experiments on all 217 countries in our datasets, as well as on the top 25 countries, we offer some insights into the best use of tweet-inherent features for an accurate country-level classification of tweets. We find that the use of a single feature, such as the use of tweet content alone – the most widely used feature in previous work – leaves much to be desired. Choosing an appropriate combination of both tweet content and metadata can actually lead to substantial improvements of between 20% and 50%. We observe that tweet content, the user's self-reported location and the user's real name, all of which are inherent in a tweet and available in a real-time scenario, are particularly useful to determine the country of origin. We also experiment on the applicability of a model trained on historical tweets to classify new tweets, finding that the choice of a particular combination of features whose utility does not fade over time can actually lead to comparable performance, avoiding the need to retrain. However, the difficulty of achieving accurate classification increases slightly for countries with multiple commonalities, especially for English and Spanish speaking countries...

Keywords

Geo, tweets, Social Media, locating, classification...

I. Introduction

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

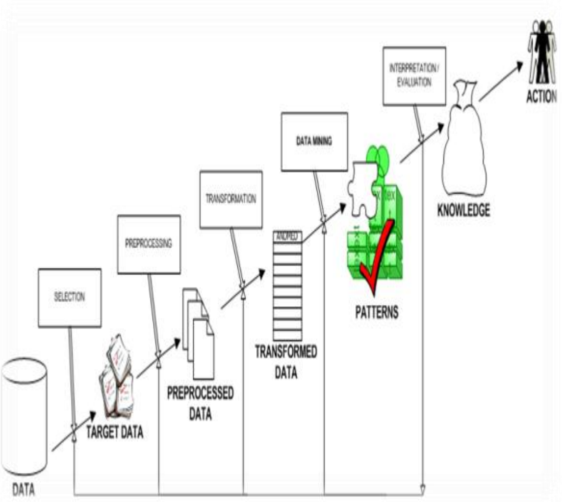


Fig 1: Structure of Data Mining

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

- Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

- Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

- Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data mining consists of five major elements:

1) Extract, transform, and load transaction data onto the data warehouse system.

2) Store and manage the data in a multidimensional database system.

3) Provide data access to business analysts and information technology professionals.

4) Analyze the data by application software.

5) Present the data in a useful format, such as a graph or table.

Different levels of analysis are available:

- Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

- Genetic algorithms: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

- Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

- Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k=1). Sometimes called the k-nearest neighbor technique.

- Rule induction: The extraction of useful if-then rules from data based on statistical significance.

- Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships

Characteristics of Data Mining:

- Large quantities of data: The volume of data so great it has to be analyzed by automated techniques e.g. satellite information, credit card transactions etc.

- Noisy, incomplete data: Imprecise data is the characteristic of all data collection.

- Complex data structure: conventional statistical analysis not possible

- Heterogeneous data stored in legacy systems

Benefits of Data Mining:

- It’s one of the most effective services that are available today. With the help of data mining, one can discover precious information about the customers and their behavior for a specific set of products and evaluate and analyze, store, mine and load data related to them

- An analytical CRM model and strategic business related decisions can be made with the help of data mining as it helps in providing a complete synopsis of customers

- An endless number of organizations have installed data mining projects and it has helped them see their own companies make an unprecedented improvement in their marketing strategies (Campaigns)

- Data mining is generally used by organizations with a solid customer focus. For its flexible nature as far as applicability is concerned is being used vehemently in applications to foresee crucial data including industry analysis and consumer buying behaviors

- Fast paced and prompt access to data along with economic processing techniques have made data mining one of the most suitable services that a company seek

Advantages of Data Mining:

- Marketing / Retail:

Data mining helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign…etc. Through the results, marketers will have appropriate approach to sell profitable products to targeted customers.

Data mining brings a lot of benefits to retail companies in the same way as marketing. Through market basket analysis, a store can have an appropriate production arrangement in a way that customers can buy frequent buying products together with pleasant. In addition, it also helps the retail companies offer certain discounts for particular products that will attract more customers.

- Finance / Banking

Data mining gives financial institutions information about loan information and credit reporting. By building a model from historical customer's data, the bank and financial institution can determine good and bad loans. In addition, data mining helps banks detect fraudulent credit card transactions to protect credit card's owner.

- Manufacturing

By applying data mining in operational engineering data, manufacturers can detect faulty equipments and determine optimal control parameters. For example semi-conductor manufacturers has a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are lot the same and some for unknown reasons even has defects. Data mining has been applying to determine the ranges of control parameters that lead to the production of golden wafer. Then those optimal control parameters are used to manufacture wafers with desired quality.

- Governments

Data mining helps government agency by digging and analyzing records of financial transaction to build patterns that can detect money laundering or criminal activities.

- Law enforcement:

Data mining can aid law enforcers in identifying criminal suspects as well as apprehending these criminals by examining trends in location, crime type, habit, and other patterns of behaviors.

- Researchers:

Data mining can assist researchers by speeding up their data analyzing process; thus, allowing those more time to work on other projects.

Objective Of The System:

A query may have multiple facets that summarize the information about the query from different perspectives. Query facets provide interesting and useful knowledge about a query and thus can be used to improve search experiences in many ways.

• First, we can display query facets together with the original search results in an appropriate way. Thus, users can understand some important aspects of a query without browsing tens of pages. For example, a user could learn different brands and categories of watches. We can also implement a faceted search [1], [2], [3], [4] based on the mined query facets.

• Second, query facets may provide direct information or instant answers that users are seeking.

• Third, query facets may also be used to improve the diversity of the ten blue links. We can re-rank search results to avoid showing the pages that are near-duplicated in query facets at the top. Query facets also contain structured knowledge covered by the query, and thus they can be used in other fields besides traditional web search, such as semantic search or entity search.

• This method is likely not to be suitable for the domain of e-commerce, where also small data sets occur and statistically deriving interesting attributes is not possible.

• Approach does not consider numeric facets and the use of disjunctive semantics for values.

• Large number of facets are available. Displaying all facets may be a solution when a small number of facets is involved, but it can overwhelm the user for larger sets of facets.

II. Literature review

A survey of location inference techniques on twitter.

The increasing popularity of the social networking service, Twitter, has made it more involved in day-to-day communications, strengthening social relationships and information dissemination. Conversations on Twitter are now being explored as indicators within early warning systems to alert of imminent natural disasters such as earthquakes and aid prompt emergency responses to crime. Producers are privileged to have limitless access to market perception from consumer comments on social media and micro blogs. Targeted advertising can be made more effective based on user profile information such as demography, interests and location. While these applications have proven beneficial, the ability to effectively infer the location of Twitter users has even more immense value. However, accurately identifying where a message originated from or an author's location remains a challenge, thus essentially driving research in that regard. In this paper, we survey a range of techniques applied to infer the location of Twitter users from inception to state of the art. We find significant improvements over time in the granularity levels and better accuracy with results driven by refinements to algorithms and inclusion of more spatial features.

Overview of replab 2013: Evaluating online reputation monitoring systems.

This paper summarizes the goals, organization, and results of the second RepLab competitive evaluation campaign for Online Reputation Management Systems (RepLab 2013). RepLab focused on the process of monitoring the reputation of companies and individuals, and asked participant systems to annotate different types of information on tweets containing the names of several companies: first tweets had to be classified as related or unrelated to the entity; relevant tweets had to be classified according to their polarity for reputation (Does the content of the tweet have positive or negative implications for the reputation of the entity?), clustered in coherent topics, and clusters had to be ranked according to their priority (potential reputation problems had to come first). The gold standard consists of more than 140,000 tweets annotated by a group of trained annotators supervised and monitored by reputation experts.

A survey of techniques for event detection in twitter.

Twitter is among the fastest-growing micro blogging and online social networking services. Messages posted on Twitter tweets have been reporting everything from daily life stories to the latest local and global news and events. Monitoring and analyzing this rich and continuous user-generated content can yield unprecedentedly valuable information, enabling users and organizations to acquire actionable knowledge. This article provides a survey of techniques for event detection from Twitter streams. These techniques aim at finding real-world occurrences that unfold over space and time. In contrast to conventional media, event detection from Twitter streams poses new challenges. Twitter streams contain large amounts of meaningless messages and polluted content, which negatively affect the detection performance. In addition, traditional text mining techniques are not suitable, because of the short length of tweets, the large number of spelling and grammatical errors, and the frequent use of informal and mixed language. Event detection techniques presented in literature address these issues by adapting techniques from various fields to the uniqueness of Twitter. This article classifies these techniques according to the event type, detection task, and detection method and discusses commonly used features. Finally, it highlights the need for public benchmarks to evaluate the performance of different detection approaches and various features.

Discriminating gender on twitter

Accurate prediction of demographic attributes from social media and other informal online content is valuable for marketing, personalization, and legal investigation. This paper describes the construction of a large, multilingual dataset labeled with gender, and investigates statistical models for determining the gender of uncharacterized Twitter users. We explore several different classifier types on this dataset. We show the degree to which classifier accuracy varies based on tweet volumes as well as when various kinds of profile metadata are included in the models. We also perform a large-scale human assessment using Amazon Mechanical Turk. Our methods significantly out-perform both baseline models and almost all humans on the same task.

Predicting twitter user locations with spatial word usage.

We study the problem of predicting home locations of Twitter users using contents of their tweet messages. Using three probability models for locations, we compare both the Gaussian Mixture Model (GMM) and the Maximum Likelihood Estimation (MLE). In addition, we propose two novel unsupervised methods based on the notions of Non-Localness and Geometric-Localness to prune noisy data from tweet messages. In the experiments, our unsupervised approach improves the baselines significantly and shows comparable results with the supervised state-of-the-art method. For 5,113 Twitter users in the test set, on average, our approach with only 250 selected local words or less is able to predict their home locations (within 100 miles) with the accuracy of 0.499, or has 509.3 miles of average error distance at best.

III. System Design and Architecture

A. *Existing System*

- The work by Han et al. is the existing system; it conducted a comprehensive study on how Twitter users can be geo located by using different features of tweets. They analyzed how location indicative words from a user’s aggregated tweets can be used to geo locates the user. However, this requires collecting a user’s history of tweets, which is not realistic in our real-time scenario.

- They also looked at how some metadata from tweets can be leveraged for classification, achieving slight improvements in performance, but again this is for a user's aggregated history.

- Finally, they looked at the temporality of tweets, using an old model to classify new tweets, finding that new tweets are more difficult to classify. This is an insightful study, which also motivates some of the settings and selection of classifiers in our own study; however, while an approach based on location indicative words may be very useful when looking at a user's aggregated tweets, it is rather limited when – as in our case – relying on a single tweet per user.

Disadvantages Of Existing System:

- 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

- Most of the previous research in inferring tweet geo location 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.

- 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.

B. Proposed System:

- Our methodology enables us to perform a thorough analysis of tweet geo location, revealing insights into the best approaches for an accurate country-level location classifier for tweets.

- We find that the use of a single feature like content, which is the most commonly used feature in previous work, does not suffice for an accurate classification of users by country and that the combination of multiple features leads to substantial improvement, outperforming the state-of-the-art real-time tweet geo location classifier; this improvement is particularly manifest when using metadata like the user's self-reported location as well as the user's real name.

- 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.

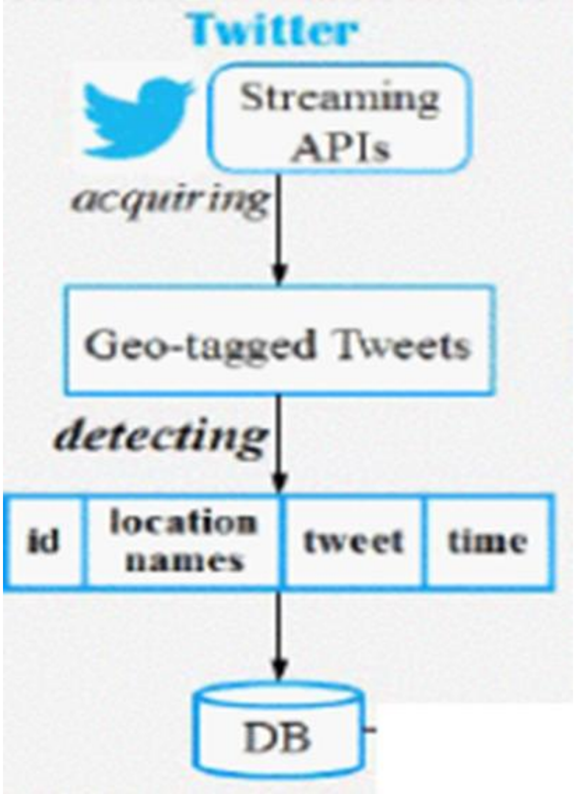


Fig 2: System Architecture

- We show that country-level classification of an unfiltered Twitter stream is challenging. It requires careful design of a classifier that uses an appropriate combination of features.

Advantages Of Proposed System:

- 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.

- We also complement previous work by investigating the extent to which a classifier trained on historical tweets can be used effectively on newly harvested tweets.

- Our results at the country level are promising enough in the case of numerous countries, encouraging further research into finer grained geo location of global tweets.

- Still, our experiments show that we can achieve F1 scores above 80% in many of these cases given the choice of an appropriate combination of features, as well as an overall performance above 80% in terms of both micro-accuracy and macro-accuracy for the top 25 countries

IV. Results

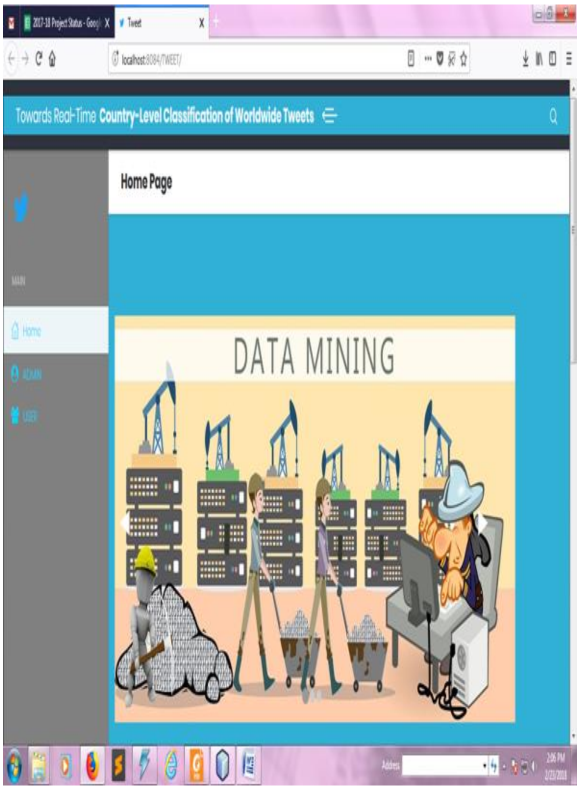

localhost:8084/TWEET/
Towards Real-Time Country-Level Classification of Worldwide Tweets
Home Page
MAIN
Home
ADMIN
USER
DATA MINING

Fig 3: Twitter

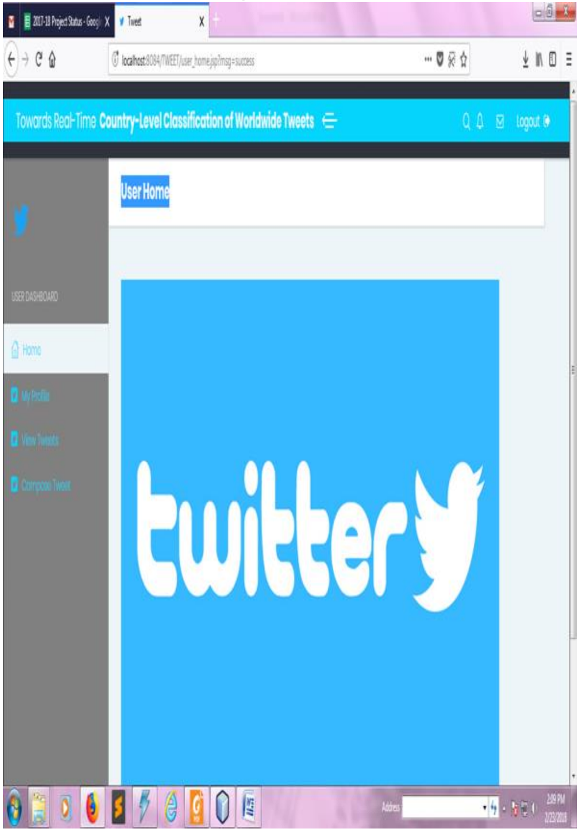
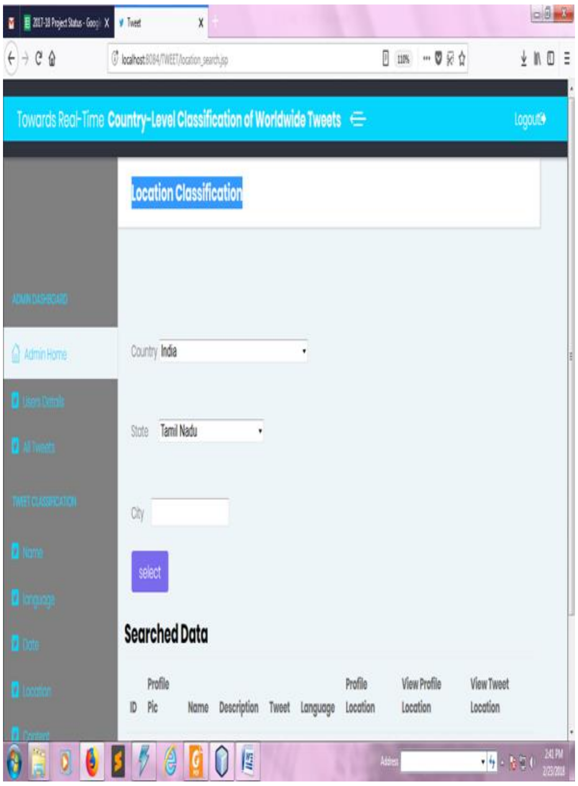

localhost:8084/TWEET/user_home.jsp?msg=success
Towards Real-Time Country-Level Classification of Worldwide Tweets
Logout
User Home
USER DASHBOARD
Home
My Profile
View Tweets
Compose Tweet
twitter

Fig.4 User Home



Location Classification

Country India

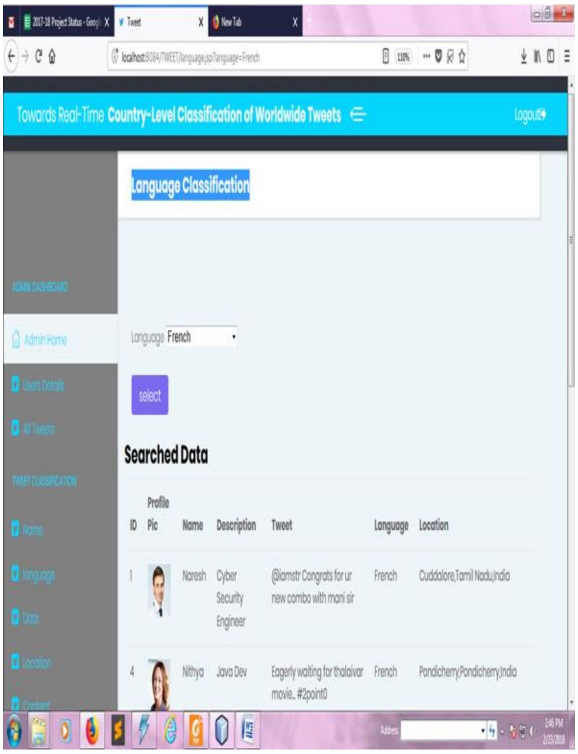
State Tamil Nadu

City

Searched Data

ID	Profile Pic	Name	Description	Tweet	Language	Profile Location	View Profile Location	View Tweet Location

Fig.5 Location Classification


localhost:8084/TWEET/language.jsp?language=French
Towards Real-Time Country-Level Classification of Worldwide Tweets
Logout
ADMIN DASHBOARD
Admin Home
Users Details
All Tweets
TWEET CLASSIFICATION
Name
language
Date
Location
Content
select

Language Classification

Language French

Searched Data

ID	Profile Pic	Name	Description	Tweet	Language	Location
1		Naresh	Cyber Security Engineer	@iamstr Congrats for ur new combo with mani sir	French	Cuddalore,Tamil Nadu,India
4		Nithya	Java Dev	Eagerly waiting for thalaivar movie.. #2point0	French	Pondicherry,Pondicherry,India

Fig.6 Language classification

V. 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 from a global stream of tweets written in any language. Most previous work focused on classifying tweets coming from a single country and 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 from a global stream of tweets written in any language.

Most previous work focused on classifying tweets coming from a single country and hence assumed that tweets from that country were already identified. Where previous work had considered tweets from all over the world, the set of features employed for the classification included features, such as a user's social network, that are not readily available within a tweet and so is not feasible in a scenario where tweets need to be classified in real-time as they are collected from the streaming API. Moreover, previous attempts to geo locate global tweets tended to restrict their collection to tweets from a list of cities, as well as to tweets in English; this means that they did not consider the entire stream, but only a set of cities, which assumes prior preprocessing.

Finally, 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 insights that can be used effectively 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 identify all the content posted from a specific country. In the future we plan to test alternative cost-sensitive learning approaches to the one used here, focusing especially on collection of more data for under-represented countries, so that the classifier can be further improved for all the countries. Furthermore, we plan to explore more sophisticated approaches for content analysis, e.g. detection of topics in content (e.g. do some countries talk more about football than others?), as well as semantic treatment of the content. We also aim to develop finer-grained classifiers that take the output of the country-level classifier as input.

VI. 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.