CLASSIFICATION OF WORLDWIDE TWEETS BASED ON COUNTRY-LEVEL LOCATION CLASSIFICATION

[1]Miss. Ruhina Syed M, [2]Dr. B. M. Patil

[1]M.Tech Students and [2]Dean of P.G.

[1]Department of P.G.

MBESs College of Engineering, Ambajogai,

Maharashtra, India.

Abstract: — Social media is a platform to express once view in a real time. This real time view of social media makes in an attractive tool for 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 geo-location has classified tweets by location within a limited geographical area or country; here we overcame the task in a broader context by classifying global tweets at the country level in a real-time scenario. The characterization of tweets is done based on utilizing geo-location. With the enthusiasm for utilizing internet-based life as the hotspot for research has inspired handling the test of naturally geo-finding tweets. To break down the tweet's nation of root this can be controlled by utilizing some tweet-inalienable highlights and Streaming LDA calculation for arrangement. Picking a fitting mix of both tweet substance and metadata can really prompt significant enhancements. We also work on finding the location of posted tweet commented by user and sentiment analysis on that comments, means we given a feedback on comment posted.

Keywords: Geo-Location, Metadata, NLP, Tweets, Twitter, Microblogging, Real-time, classification, etc.

I. INTRODUCTION

Today social media like YouTube, WhatsApp, Facebook, Hike, Twitter etc. are rapidly being used in the different sector like 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. A public API that also useful for different purpose that allow user can use it free of cost collection of large amount of data, Twitter has become a best data source for such s-learning. That new kind of data score twitter provided, a person who research working on a development tools for real-time analytics as well as into sentiment analytical approaches for understanding the post expressed by users towards a public or target user opinion on a specific tweet [5]. 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. Automated inference of social media demographics would be useful, among others, to broaden demographically aware social media analyses that are conducted through surveys [16]. In those demographic details there was a missing part is user country of origin, that issue we study here. The only option then for the researcher is to try2 meant towards a topic broken down by country. To the best of our knowledge, our work is the first to deal with global tweets in English language, using only those features present within the content of a tweet and its associated metadata.

We motivated by the previous work because in existing work they were developed an application that indicate only a particular location of tweet or specific location or limited location of tweet, in our work we develop an application to identify a geo-location tweet 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 [14], we explore the usefulness of some tweet-inherent features, all of which are readily available from a tweet object as retrieved from the Twitter API, for determining its geo-location. 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. 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. 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. 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. 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. Cases where country level geo-location is more challenging include English and Spanish speaking countries, which are harder to distinguish due to their numerous commonalities. 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.

II.LITERATURE SURVEY

The rapidly increase of interest using in social media considering since last few year researchers working on the different social media, Arkaitz Zubiaga1, Alex Voss2 et.al, has worked on a social media considering twitter to finding their tweet post location of user. In that they used the maximum entropy classification algorithm and only find out the location of country of posted tweet here they were also worked on consideration of two datasets and finding the accuracy of that. In their module, using eight different features they find the country location of different user. O. Ajao, J. Hong et.al worked on twitter for survey of location inference technique, in that they finding the location of user in a limited geo-graphical area or we can say that a one city or one country. In that they used the state-of-the art techniques was used. S. Hemamalini, K. Kannan et.al in this they were used a k-means technology and finding location based on friends and followers.

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 a tweet posted from anywhere in the world. We look at the suitability of some tweet features for this purpose and experiment on one datasets collected within different time frames to measure the usefulness. Also we work on sentiment analysis on comment posted by different user.

III. METHODOLOGY

The main workflow used Twitter as a data set source, but Twitter could handily be replaced by other streamable sources. This workflow was further divided into four distinct phases, as illustrated in Fig.3.1. Each phase is detailed in a section below. Also in this section we discuss about our system framework diagram with their explanation in details and the extra features that we added in that proposed system related to feedback as follows.

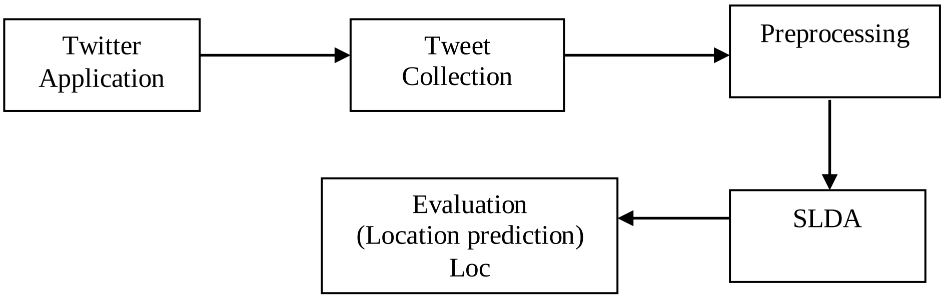

Fig.3.1 Methodology workflow

- **Tweet collection**

Twitter provides access to its documents (tweets) by means of the Twitter API [39]. Apache Flume, which uses this API, was used to collect certain sets of tweets. It could readily stream large amounts of tweets and store these onto an Apache Hadoop-compliant distributed resilient storage. Twitter was also used in a Java implementation to directly access the Twitter API and download tweets from specific users.

- **Pre-Processing**

For experimental purposes a number of different subsets were created. For example, data sets tended to have very large amounts of re-tweets, and so sub selections were made where re-tweets were either eliminated completely, or heavily reduced. These sub-selections are detailed below:

  - Re-tweets were either eliminated completely, or reduced in count to of the number of re-tweets N of the original tweet.
  - To keep the document, count tractable, sampled selections of tweets could be taken from tweet sets; for example, randomly (uniformly) choosing 10% of the documents.

The sets of tweets were accessed using a Java program developed to use Apache Spark [40]. Spark is a distributed/cluster computing solution with very useful features for resilience, fault-tolerance, integration with Hadoop, and inherent parallelism.

- **SLDA**

First, a streaming source of tweets is specified, which may either be an offline set of possibly preprocessed tweets stored with Hadoop, or a live stream of tweets (in which case preprocessing happens live). SLDA parameters (α, β, K etc.) are also specified here. Our SLDA model admits this assumption. In each iteration, the correct, corpus-wide (global) φ is copied to all worker nodes. z is then resampling locally, updating φ on each node as if all other nodes were idle. At the end of the iteration, when all nodes have resampled the topic assignment z of each word in each document, a reduce operation is used to add up all nodes' separate counts and compute the correct global φ again. Then the next iteration starts. These iterations are repeated according to the Gibbs iteration parameter, allowing z to "converge."

At the end, φ and θ for the window in time t are computed and output to file annotated with the window's date. θ is saved via Hadoop, much like the tweets are stored. Each vocabulary is also saved to file. The model (defined by z) is not immediately discarded afterwards; it is used to calculate the prior of the next model. Ultimately, these θt and φt can then be used to discover trends and topics, and perform an evaluation of the model's performance.

- **Evaluation**

Location Prediction When SLDA completes, θ is inferred for each tweet in the training set. The assumption θ ~DirK(α) can then be used to infer topic proportions for any tweet, allowing the model to make testable predictions. For implicit evaluation, θ is inferred for tweets in the sample set. The model, for some window, then used θ · φ to predict the most likely location word in the tweet. If the predicted location matches the election location at some time, that prediction is marked as correct. The procedure is similar for explicit evaluation, except that θ is 1 for the human-chosen topic and 0 elsewhere. The confidences of these predictions, normalized over the 10 most likely locations, were also recorded.

IV. SYSTEM ARCHITECTURE

In Fig 4.1 the system model on our scheme includes two modules.

- **User:**

In User module, Initially User must have to **register** their detail (First name, last name, Date of birth, Gender, Email-id, Mobile Number, Username and Password all information filling is must) and after registration user can login using the user name and their password.

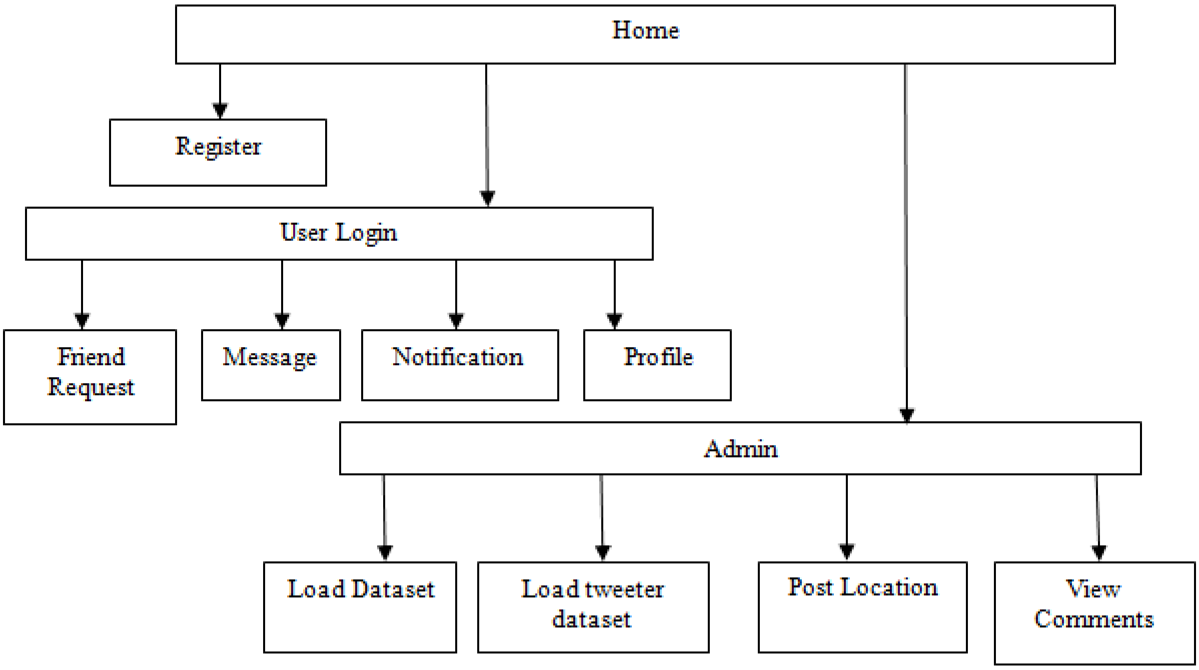

Fig 4.1 System Architecture

After login user can view four options as-

- **Friend Request:**

In this user can send a request through a message to her/him particular friend and user also see the friend list in friend request part.

- **Message:**

In this user can post the message to friends and also see the comments posted by their friends. User can also post an image to friends and comment on that. We also added one extra feature feedback based upon the user comment.

- **Notifications:**

Here user get notification related to post of their friends and their friend's comments on his/her post.

- **Profile:**

The user sees their profile details what they entered at registration time, contact details. User can change their password also and contact details for example mobile number and address. User also uploads a photo on their account profile

- **Admin:**

In Admin module, Admin can view all the user's details in the database with their location details. Admin can search the tweets using specific user id of post and that gives all the information of tweet. It also has a sub parts

- **Load dataset and load tweet dataset**

Here we make a dataset that we can loaded

- **Post Location**

When user post a tweet to another user than that user current location is shows

- **View comments**

  Here we can see the entire tweet posted information in details with user id, Post id, First name, Last name, Post comment, and map. Clicking on map pointer icon we see the posted comment location.

In this work we also implemented that, when user posted a tweet and their friends comment on that then we find out which type of comment it is positive or negative.

This feedback we finding out using two functions as follows

Positivefeedback()

Negativefeedback()

In this way we implement our proposed system

V. RESULT ANALYSIS

In fig.4.8 we show the system performance comparing with existing and proposed model as

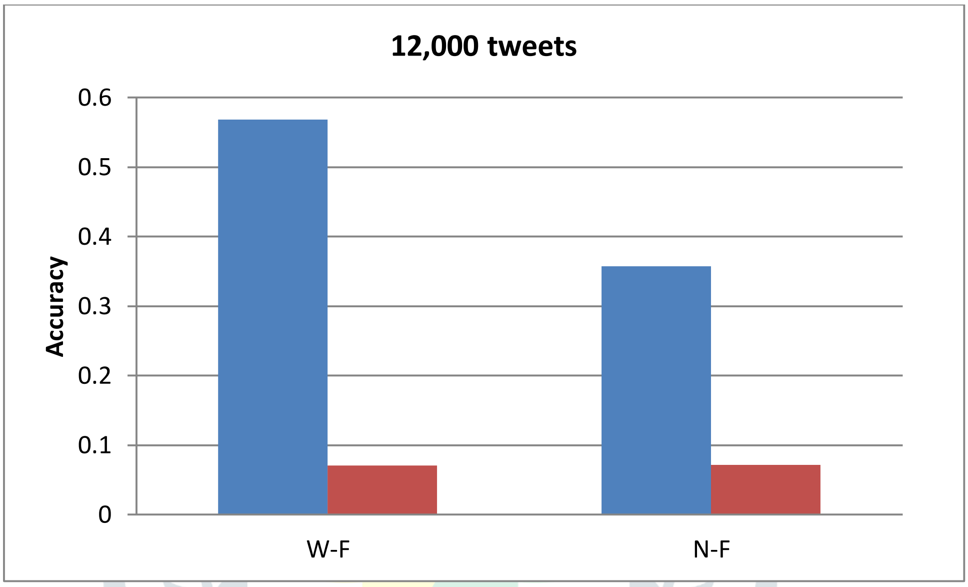

Fig.5.1 System Performance

The above fig.4.8 shows the system performance of existing system and proposed system of accuracy level, we observe that existing system having a lager accuracy then existing one.

12,000 Tweets		
Techniques	W-F	N-F
Existing System	0.568	0.3568
Proposed System	0.0703	0.0713

Table 5.1 System Performance

Table 5.1 shows the system performance in tabular form. So we can get clearly idea of accuracy of existing system and proposed system.

VI. CONCLUSION

To the simplest of my data, the ubiquity and increasing capability of social media have made it a growing target for study and extraction of useful information, Here, we prototypical tool is used with which the topics and trending locations of streaming media can be automatically discovered. The tool provides key abilities to users to gain an overview of online discussions and features to search, this is the 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 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. In the work presented here in, the streaming latent Dirichlet allocation (SLDA) topic model has been evaluated to be an effective model for tracking geographical trends and topical locations. 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 desire to spot all the content from a particular country.

VII. FUTURE WORK

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 plan to hack the location or we can say that we build a model in that we can detect a wrong location to posted tweet user. We also plan to change the user location of specific user for example, if valid user is a terrorist and they want to send a post and they send a post or comment on twitter at that time there location is captured in proposed system in future work that terrorist want to change their wrong location means he posted from any town in china any indicate the location of that person in any town in Pakistan, in that way we also implement that .Also User can used these techniques for giving a valid information of fraud person but posted person don't want to show their current location for their security purpose. We also aim to develop finer-grained classifiers that take the output of the country-level classifier as input.

VIII. REFERENCE

[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, IJCSET pages 509-513,2015

[15] Elsevier B.V. Tracking geographical locations using a geo-aware topic model for analyzing social media data. This is an open access article under the CC BY license, 2017.

[16] Hesam Amoualian, Eric Gaussier Streaming-LDA: A Copula-based Approach to Modeling Topic Dependencies in Document Streams KDD '16, August 13-17, 2016, San Francisco, CA, US

[17] M. Graham, S. A. Hale, and D. Gaffney. Where in the world are you? Geo-location and language identification in twitter. The Professional Geographer, 66(4):568–578, 2014.

[18] B. Han, P. Cook, and T. Baldwin. Text-based twitter user geolocation prediction. Journal of Artificial Intelligence Research, pages 451–500, 2014.

[19] B. Hecht, L. Hong, B. Suh, and E. H. Chi. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In Proceedings of CHI, pages 237–246, 2011.

[20] B. R. Heravi and I. Salawdeh. Tweet location detection. In Computation + Journalism Symposium, 2015