Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 03, March -2018

TOWARDS REAL TIME,COUNTRY LEVEL LOCATION CLASSIFICATION OF WORLDWIDE- FACEBOOK

Arockia Selvaraj.A[1],Dhinesh Kumar.S[2],Dhivaharan.S[3],Prakash.A[4],SureshKumar.S[5]

[1]*Assistant Professor, Info Institute of Engineering, Kovilpalayam, Coimbatore- 641 107*
[2,3,4,5]*UG graduate, Info Institute of Engineering, Kovilpalayam, Coimbatore- 641 107*

Abstract: *The increase of interest in using social media as a source for research has motivated tackling the challenge of automatically geolocating messages , given the lack of explicit location information in the majority of messages. In contrast to much previous work that has focused on location classification of messages restricted to a specific country, here we undertake the task in a broader context by classifying global messages at the country level, which is so far unexplored in a real-time scenario. We analyse 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 analyse the extent to which a model trained from historical messages can still be leveraged for classification of new messages. 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 messages. 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 messages to classify new messages, 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: *User Location, TimeZone, Message Classification, Country Classification, GeoLocation.*

I INTRODUCTION

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 . 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. 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. 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. The classifier built on this pre-filtered dataset may not be applicable to a Twitter stream where every tweet needs to be geolocated. 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. 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. 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.

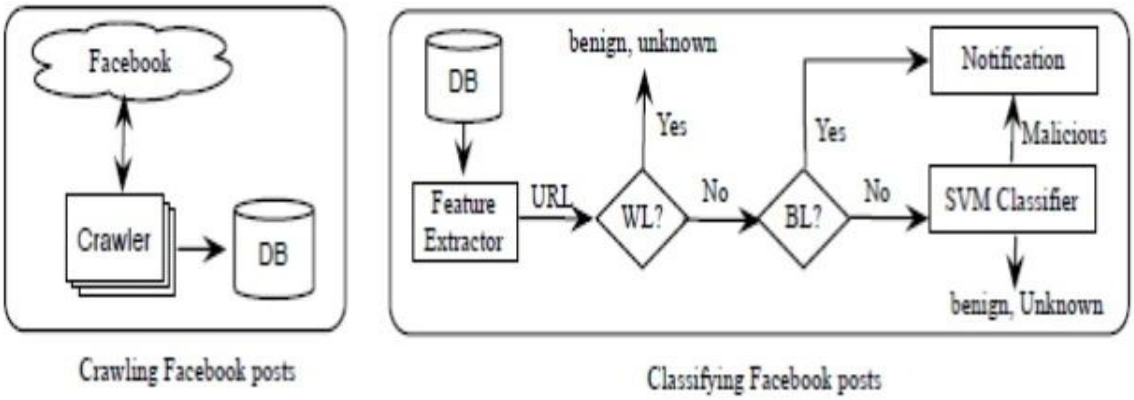


Figure 1-System Architecture

II EXISTING SYSTEM

In this work 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 messages to classify new messages, finding that the choice of a particular combination of features whose utility does not fade overtime can actually lead to comparable performance, avoiding the needto retrain Disadvantages: There is no big difference between the two approaches based on GeoNames when we look at micro-accuracy Note that while higher values are desired for micro-accuracy and macro-accuracy, lower values are optimal for MSE.

III 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.

IV 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 theclassifier.

V SOFTWARE IMPLEMENTATION

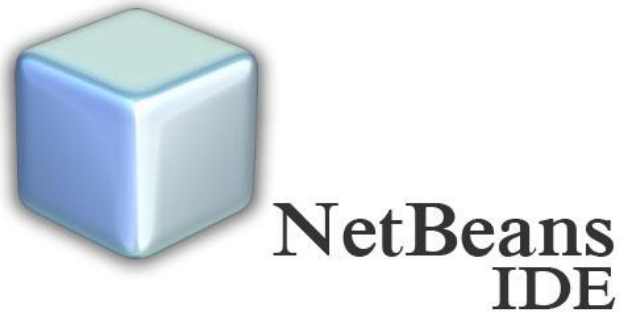


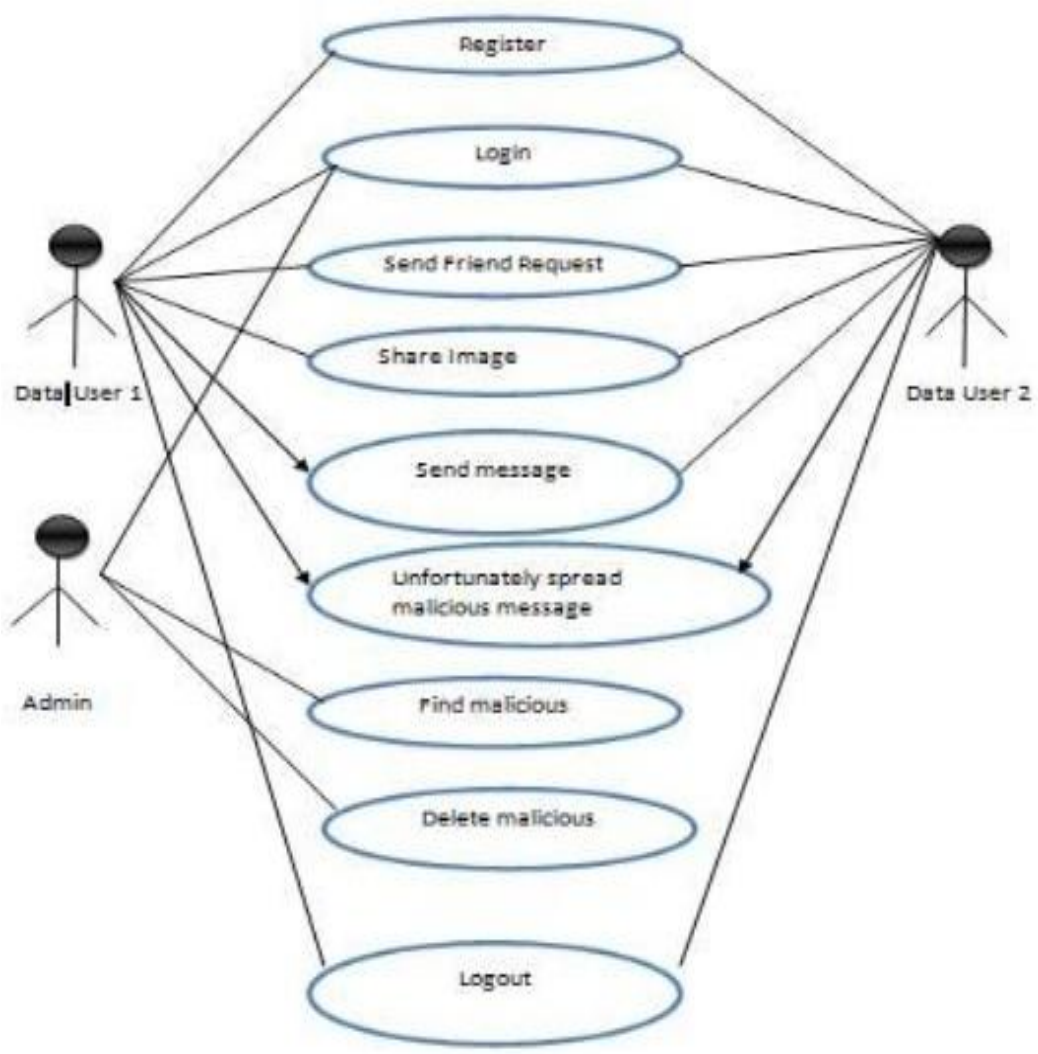
Figure 6- Netbeans IDE

5.1. DESCRIPTION

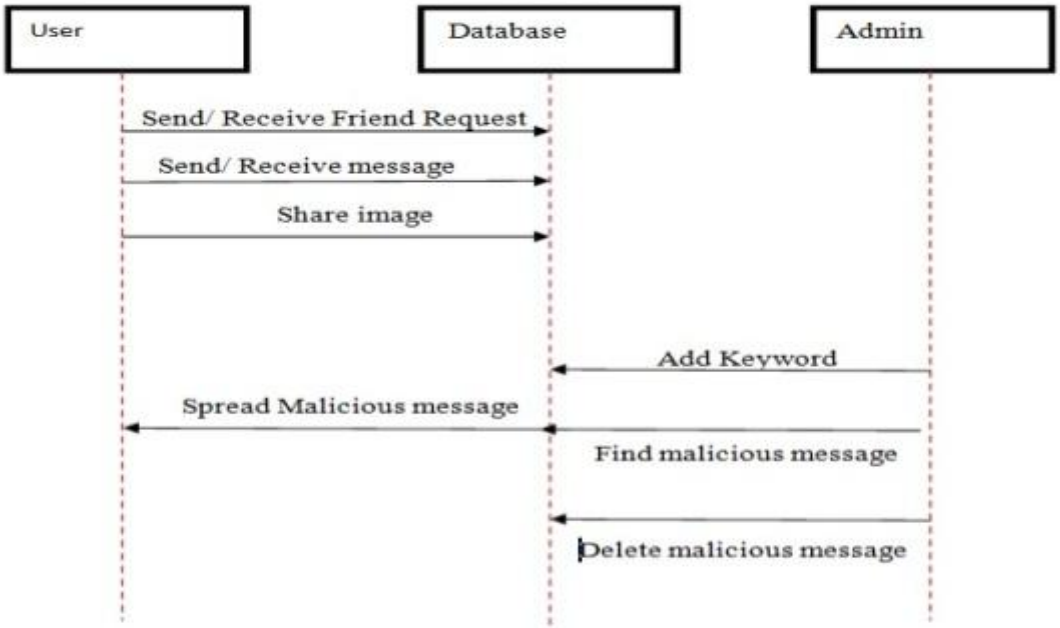
NetBeans IDE lets you quickly and easily develop Java desktop, mobile, and web applications, as well as HTML5 applications with HTML, JavaScript, and CSS. The IDE also provides a great set of tools for PHP and C/C++ developers. It is free and open source and has a large community of users and developers around the world.

VI UML DIAGRAMS

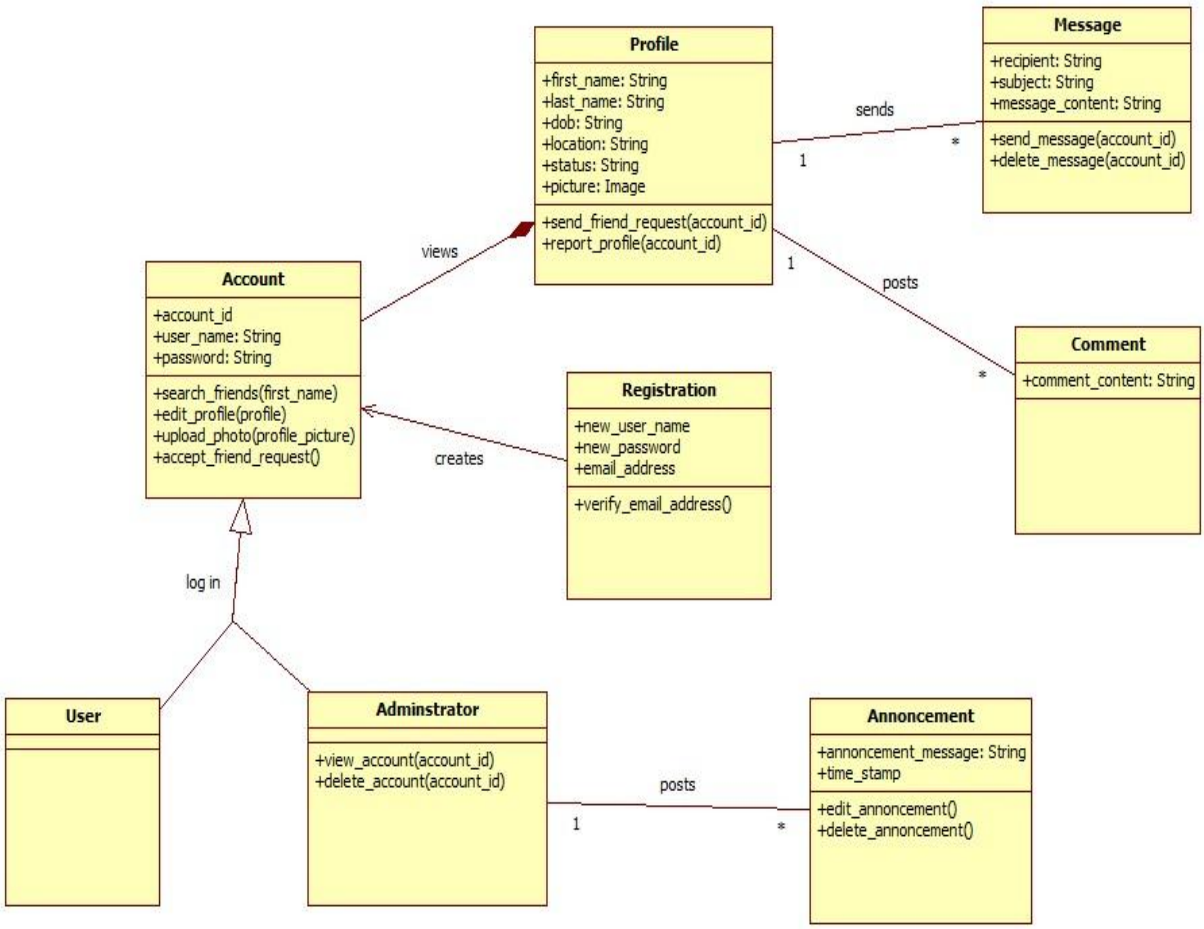
6.1. USECASE DIAGRAM



6.2. SEQUENCE DIAGRAM



6.3. USE CASE DIAGRAM



VII ENVIRONMENT

An integrated development environment (**IDE**) is a software application that provides comprehensive facilities to computer programmers for software development. An**IDE** normally consists of a source code editor, build automation tools, and a debugger. Most modern **IDEs** have intelligent code completion**.** IDE is based on Processing Programming language and supports ARDUINO language. **IDE** (Integrated Drive Electronics) is a standard electronic interface used between a computer motherboard's data paths or bus and the computer's **disk** storage devices.Usually, IDE works by connecting to a computer through an Integrated Drive Electronics (**IDE**) interface. Essentially, an **IDE** interface is a standard way for a storage device to connect to a computer

VIII 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 messages in a real-time scenario from a global stream of messages written in any language. Most previous work focused on classifying messages coming from a single country and hence assumed that messages from that country were already identified. Where previous work had considered messages 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 withina tweet and so is not feasible in a scenario where messages need to beclassified in real-time as they are collected from the streamingAPI.

IX 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.

