Identifying System Location Specifics based on Classification of Worldwide Tweets

Aatmakuri Hima, Sirasanagondla Venkata Naga Sreenivasu

***Abstract:** As social networking sites are gaining populism across the globe, people are more enthusiastic about sharing their thoughts on Various networking Platforms. Facebook and Twitter have become a leading destination for sharing various kinds of information. In the existing literature the focus is to access the information published in the networking platforms in the real-time, and they do not focus on obtaining the geo-location of the user. Here we propose a monitoring system that classifies the tweets using some reliable techniques which can be used across the globe without any security concerns. As there is a lot of fake news available in the digital form, there is a definite need to access the user information and his geo-location metrics. In this paper, we have introduced Naive Bayes Multinomial classifier and a few other models which performs a spatiotemporal analysis. This study also identifies a comprehensive set of performance metrics which can access the tweet's country of origin by using eight tweet-inherent features. The outcome of this analysis can be used by various cyber-crime departments to deal with the numerous cybercrime cases on networking platforms, and real-time decisive actions can be taken.*

***Index Terms:** Face book, geo-location, social networking, Real-Time, Classification, Tweets country of origin, Message Classification.*

I. INTRODUCTION

Usage of social networks is increasing tremendously, with the use of Facebook, Twitter, Instagram, etc. Users companionate their feelings, thoughts, ideas, and interests in these sites. However, in most of the sites, we don't have access to the user's information, and hence there is lack of clear-cut location details in most of the tweets. Twitter is not only used for chatting purpose it can also be used for sharing the posts and give their comments about any posts, and they can also share their emotions, even health conditions, etc. [3]. Various machine learning algorithms such as Naive Bayes and other ensemble approaches are implemented to classify the terrorism data [4]. This methodology carries a critical assessment of state of the fine art by testing nine geolocation inference techniques, and all published recently in top-tier conferences. We release an open-source geo inference framework that includes implementations off all of the nine methods and proposed evaluation metrics, lowering the bar for future comparisons [5]. But in this project, we provide a user's location information which can access all countries. There is a difficulty in making classification of some countries which is having multiple similarities, especially who are using Spanish and English language.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author

Aatmakuri Hima*, pursuing M.Tech. (CSE), Narasaraopet Engineering College, Narasaraopeta, AP – 522601, India.

Dr.S.V.Naga Srinivasu, Professor, Computer science and engineering, Narasaraopeta Engineering College, Narasaraopet, AP -522601, India.

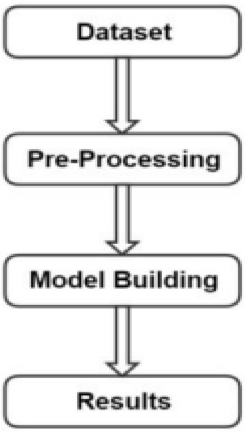


Figure 1. The flow of the prediction system

II. TECHNIQUES

In this project, we are using some algorithms. They are k-Nearest and Naive Bayes classifiers. Using this algorithm, we can get the clear-cut idea of proposed techniques, and we also do data set classification using above algorithms.

A. Naive Bayes Classifier

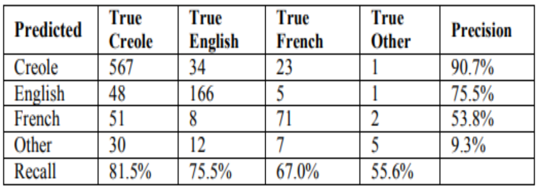
Sentiment and Language are some of the methods to predict the location of the user. Naive Bayes can predict accurately for positive sentiment. The Recall and the precision values in the negative sentiment case are quite low. The possible reason is that we have very less database for testing compared to others. The accuracy for language classification is 84% from Naive Bayes classifier. This classifier is the best in the classification of language rather than the sentiments. Prediction of English and creole is very high (89%) whereas the prediction of the French is lower [10].

Table 1. Percentage Accuracy For Sentiment Classification

Predicted	True Negative	True Positive	True Very Negative	True Neutral	Precision
Negative	228	67	41	62	57.3%
Positive	57	131	13	55	51.2%
Very Negative	48	39	31	83	15.4%
Neutral	29	63	8	76	43.2%
Recall	63.0%	43.7%	33.3%	27.5%	

B.K-Nearest Neighbor Classifier

Sentiment and language are two types of classification. Sentiment categorization is done using naive Bayes, whereas the language categorization is done by k-nearest. For both the techniques this project will predict the location more accurately. After comparing the conclusions of both algorithms, it is very conclusive that the K-Nearest gives the best results compared to any other techniques. Language categorization method accuracy was found to be 78%. Hence the accuracy of the Naive Bayes algorithm is better compared to a k-NN classifier as shown in the Table [2].

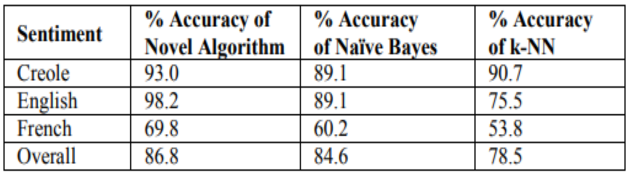


Predicted	True Creole	True English	True French	True Other	Precision
Creole	567	34	23	1	90.7%
English	48	166	5	1	75.5%
French	51	8	71	2	53.8%
Other	30	12	7	5	9.3%
Recall	81.5%	75.5%	67.0%	55.6%	

Table 2. Percentage Accuracy For Language Classification

C. Adaptive Algorithm Vs. Machine Learning:

In this section, the results of the adaptive algorithm were compared with Machine learning algorithms. The comparison results are given in the below Table. Comparison between the precision values is made as a performance metric. Hence, we can conclude that this algorithm performance is the best in all three categories. Here we propose a polarity assignment technique, and we can see that the performance of it is far superior compared to the machine learning and naive based frequency sentiment classifiers. We propose that by embedding AI and machine learning technology in the above techniques the overall accuracies of the system can be further improved.



Sentiment	% Accuracy of Novel Algorithm	% Accuracy of Naïve Bayes	% Accuracy of k-NN
Creole	93.0	89.1	90.7
English	98.2	89.1	75.5
French	69.8	60.2	53.8
Overall	86.8	84.6	78.5

Table 3. Percentage Precision For Language Classification

III. Architecture

In previous papers, they mainly concentrated only on gathering user's information without the location details of users. But in this study, it will access all country people location details there is no restriction to any country. Basically, there are 217 countries where we are using the topmost 25 countries in this project. The use of providing full details of user is to reduce cyber-attacks where we can quickly identify the person because we have the flexibility to give the exact location information. There are many people who addicted to social websites as they want to increase their popularity and some other reasons. But with this algorithm, we can save all the data of the user using twitter, and we will do a classification according to all countries. Finally, Streaming APIs obtain geotagged tweets and gathers people details like name, id, time and location details are stored in database

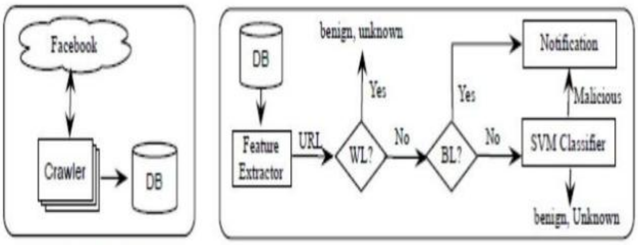


Figure 2. System Architecture

IV. RESULTS

This study on geo-located tweets from users reveals insight into their geo-location of tweets in a real-time. Testing this study with multiple features, we found that the geo-location accessing performance is substantially improved compared with the existing literature. The use of geolocation of the user to build an extensive set of data with the location located should be widely accepted. This will be helpful for the non-geolocated tweets having similar characteristics to use this methodology.

Figure 3. Training times on the one-month training data set

V.CONCLUSION AND FUTURE WORK

Based on the existing literature, this is the first time that a comprehensive study is performed to access the origin country of the tweet in the real-time of any language. Most of the existing literature is able to access the origin of the tweet only to specific countries and also for particular languages alone. Moreover, all the current research tries to access global tweets only limited to a specific number of cities and limited to English language alone. Here we apply all our algorithms on the 2 sections of data which are collected from the Internet Achieve database. So, we are utilizing all our algorithms on the existing data, validate the same and then use our study in the present social networking era. This can help the entire globe to be able to identify the new one with the algorithm validated on the old tweets. Our study can be further improved by developing many more cost-sensitive algorithms which can be used in all the developed as well as developing countries. These methods include analysing the content and then classify the tweet. For example, few countries talk more about cricket than football, etc. All these algorithms embedded with Artificial intelligence which is a very powerful tool in the 21^{st} century can help to access the geolocation of the tweet more accurately.

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. Lexpress.mu. "Cyberbullying: Akash Callikan Porte plainte a la cybercrime unit," Lexpress.mu, Mar. 16, 2014.
3. Hibz, Y. D. "Online Racial Hatred Incitement: Police Elaborated a List of 30 Suspects," Island Crisis, Sept. 8, 2015.
4. Lexpress.mu. "Cybercrime Unit: Dans l'univers des enquêteurs du virtual," Lexpress.mu, Oct. 10, 2015.
5. Dinakar, K., Reichart, R. and Lieberman, H., "Modeling the Detection of Textual Cyberbullying," in The Social Mobile Web (ICWSM) Workshop, Barcelona, Spain, 2011, pp. 11-17.

6. Gerber, M. S., “Predicting Crime Using Twitter and Kernel Density Estimation,” Decision Support Systems, vol. 61, pp. 115-125, 2014.
7. Bolla, R. A., “Crime pattern detection using online social media,” M.S. thesis. Missouri University of Science and Technology, USA, 2014.
8. Lin, Y. R., “The Ripples of Fear, Comfort and Community Identity During the Boston Bombings,” in iConference, Pittsburgh, USA, 2014, pp. 708-720.
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 192–1 99, 2011.
13. D. Doran, S. Gokhale, and A. Dagnino. Accurate local estimation of geocoordinates for social media posts. arXiv preprint arXiv:1410.4616, 2014.

AUTHORS PROFILE

Ms. Aatmakuri Hima was born in Narasaraopet, India, in 1996. She is pursuing her master’s in computer science at NEC College of Engineering, Narasaraopet, India. Her expertise lies in the field of Database management systems and Computer Networks.

Dr. Sirasanagondla Venkata Naga Sreenivasu was born in 1975. He received his Ph.D. from Acharya Nagarjuna University, Guntur, India in the field of software testing from the Department of Computer Science. He is into the teaching field from past 19 years and served various educational institutions. Being a member of IEEE, he is currently working as a Professor in Narasaraopet College of engineering, Narasaraopet, Guntur, India. He has 2 patents and more than 58 articles in reputed journals and which includes Scopus and other indexed bodies.