International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-10, Issue-4, July 2022
https://doi.org/10.55524/ijircst.2022.10.4.26
Article ID IRPP1044, Pages 213-217
www.ijircst.org

Precinct Vaticinator on Social-Media using Machine Learning Techniques

M. Chaitanya Bharathi[1], Dr. A. Seshagiri Rao[2], B. Sravani[3], and R. Veeranjaneyulu[4]

[1]Assistant Professor, Department of Information Technology, PACE Institute of Technology and Sciences, Ongole, Andhra Pradesh, India
[2]Professor, Department of Information Technology, PACE Institute of Technology and Sciences, Ongole, Andhra Pradesh India
[3]Assistant Professor, Department of Information Technology, PACE Institute of Technology and Sciences, Ongole, Andhra Pradesh, India
[4]Professor, Department of Computer Science & Engineering, PACE Institute of Technology and Sciences, Ongole, Andhra Pradesh, India

Correspondence should be addressed to M. Chaitanya Bharathi; ithod@pace.ac.in

ABSTRACT— Precinct vaticinator of users from online social media brings considerable research these days. Automatic recognition of precinct related with or referenced in records has been investigated for decades. As a standout amongst the online social network organization, Social-Media has pulled in an extensive number of users who send a millions of tweets on regular schedule. Because of the worldwide inclusion of its users and continuous tweets, precinct vaticinator on Social-Media has increased noteworthy consideration in these days. Tweets, the short and noisy and rich natured texts bring many challenges in research area for researchers. In proposed framework, a general picture of precinct vaticinator using tweets is studied. In particular, tweet precinct is predicted from tweet contents. By outlining tweet content and contexts, it is fundamentally featured that how the issues rely upon these text inputs. In this work, we predict the precinct of user from the tweet text exploiting machine learning techniques namely naïve bayes, Support Vector Machine and Decision Tree.

KEYWORDS— Social-media, Social-Media, Tweets, precinct vaticinator, Naive-Bayes, Support-Vector-Machine, Decision -Tree, Machine- Learning

I. INTRODUCTION

Users may post their precinct on the following image (see figure 1 to 3).

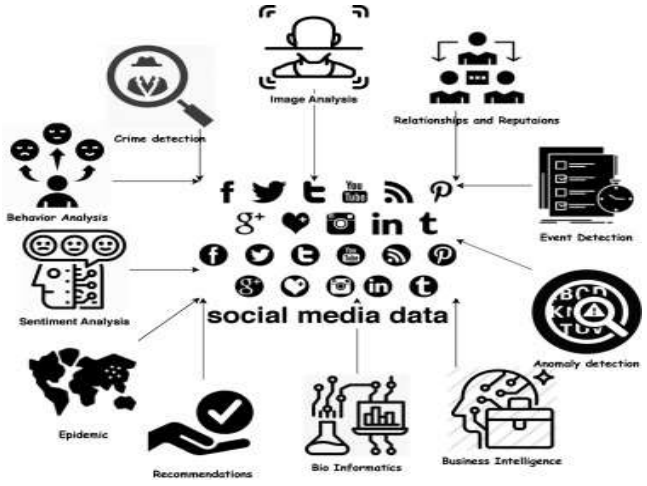

Figure 1: Home Precinct1

Tweet Precinct: Mentioned Precinct:

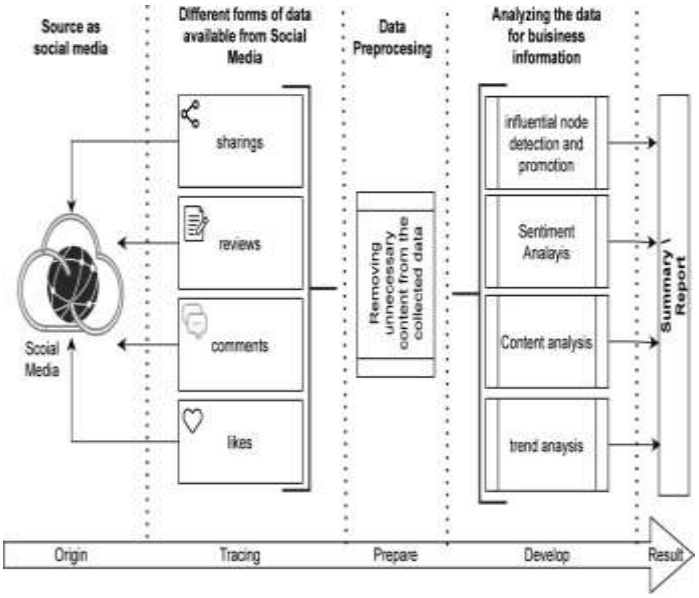

Figure 2: Home Precinct2

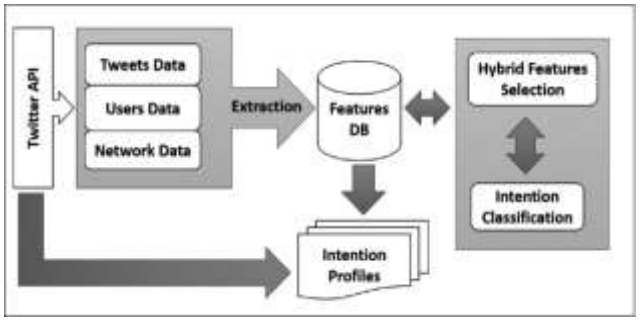

Figure 3: Block Diagram of Data Transmission

II. RELATED WORK

So many existing techniques have been studied by the researchers on precinct vaticinator problem from tweet content and social media content[1]

III. PROPOSED METHODOLOGY

Live stream of Social-Media data is collected as dataset using authentication keys. The aim of proposed system is to predict the user precinct from Social-Media content considering user home precinct, tweet precinct and tweet content. To handle this we used three machine learning approaches to make vaticinator easier and finding the best model amongst them. Fig. 4, represents the overall architecture of the proposed system with methodology modules represented.

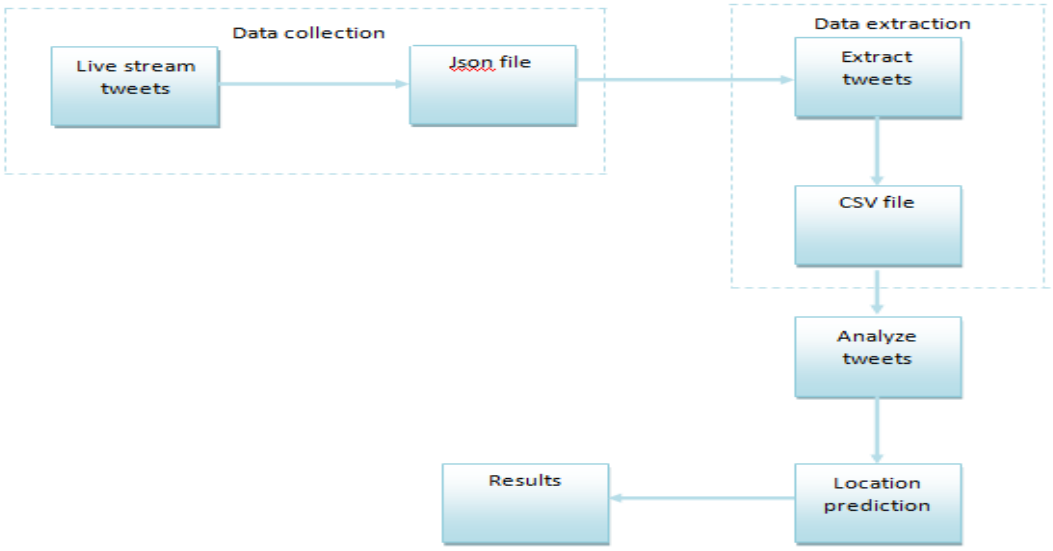

Figure 4: System Architecture

Live tweet stream from Social-Media for keyword “apple” is collected and stored in 'Social-Media.json’ file. Live Social-Media data can be collected by registering a consumer_key, consumer_secret,[2] access_token, access_token_secret for authentication and collecting live stream of tweets. We have collected more than 1000 tweets of particular keywords such as ‘Chennai, Mumbai, Kerala’.

The information extracted from live includes tweetid, name, screen_name, tweet_text, HomePrecinct, TweetPrecinct, MentionedPrecinct, Lvalue.[4]
Primary analysis was a basic processing of the text of the tweets. This was done by merging the collected tweets for a given user into a single “document” and analysing that.

A	B	C	D	E	F	G
tweet_id	Name	screen_name	tweet_text	Home Location	Tweet Location	Mention Loca Lva
7.89858E+17	Savishkar Live	SavishkarLive	RT Kerala Govt invites applications from SE	Bhopal India	Bhopal India	Kerala
1.03437E+18	cheeks	unikIIn	ito pa	puso mo	puso mo	Nil
2895688958	A Masked Error	BumchikSeenu	RT Smt Vyjayanthimala age 86 who was the	Chennai	Chennai	Nil
169426623	Jai Hind	arbind1982	RT Railway 7 2 5	Lagos Nigeria	Lagos Nigeria	Nil
8.18519E+17	Vijith	vijithfilmlover	Smt Vyjayantimala Age 86	India	India	Nil
7.43735E+17	Johns	CricCrazyJohns	RT Melbourne or Mumbai MCG crowd abou	Kerala India	Kerala India	Nil
108272890	Mahesh Veeramal	MahesMaddy	No words	Bengaluru India	Bengaluru India	Nil
3032998642	M La	ItzMiLu	RT rt Bumping into you made my day Live A	Bharat	Bharat	Nil
131520960	Ashish Chandorkar	AshishChandMT	Dear Sir This is MahishmatiThali in Pune pr	Pune	Pune	Nil
173786873	Jai Italy Jai Jai Italy	ravi enigma	some serious mental issues out there in K	Uttara Prachand	Uttara Prachand	Kerala
8.39501E+17	Austinne	Austianoo7	RT Telugu Sarkar gross gt Gang S3 Tamil Sar	Kerala India	Kerala India	Kerala
9.25043E+17	Arul Vignesh	ArulVignesh7	RT Adopted Son Of Kerala Suriya Fan Girl o	Chennai India	Chennai India	Kerala
298368845	Nelson Ji	Nelson Ji	RT FC Trade Updates Viswasam Chennai Ci	Chennai India	Chennai India	Chennai

Figure 5: Extract Live Precinct Live Social-Media

A. Data Collection and Extraction

Live tweet stream from Social-Media for keyword “apple” is collected and stored in 'Social-Media.json’ file. Live Social-Media (see figure 5) data can be collected by registering a consumer_key, consumer_secret[5], access_token, access_token_secret for authentication and collecting live stream of tweets. We have collected more than 1000 tweets of particular keyword such as ‘Chennai, Mumbai and Kerala’. The information extracted from live includes tweetid, name, screen_name, tweet_text, HomePrecinct, TweetPrecinct, MentionedPrecinct, Lvalue. Data from 'Social-Media.json’ file is read and extracted tweetid, name, screen_name, tweet_text, HomePrecinct, TweetPrecinct, MentionedPrecinct are extracted. Tweet text is compared with natural language tool kit package available in python to extract data from json file to csv is done here.

B. Data Preprocessing

Data pre-processing include the following steps,

- Extra characters are removed from tweet text.
- Capitalize all words to find for geo precinct
- Remove the tweet if user home precinct not mentioned
- Mention home precinct in tweet precinct, if user tweet precinct is null
- Removes tweets if no precinct is mentioned in tweet text.

Final extract geodata from tweet text. Last step is to assign integer value to the precincts, for example Chennai—1, Mumbai—2, Kerala—3. Lcoder is used to assign precinct as integer value.[6]
The work is implemented using Python programming, with few libraries used are scikit learn, numpy, pandas, matplotlib, geography.[8]

C. Naive-Bayes Classification

Naive-Bayes classifier is the most popular and simple classifier model used commonly. This model finds the posterior probability based on word distribution in the document. Naïve Bayes classifier work [10] with Bag Of Words (BOW) feature extraction model, which do not consider the position of word inside the document. This model used Bayes Theorem for vaticinator of particular label from the given feature set. The dataset is split into trainset and test set. Upon test set, NB_model is applied to find the precinct vaticinator[9].

D. Support Vector Machine

Support vector machine is one of most common used supervised learning techniques, which is commonly used for both classification and regression problems. The Pseudo-code works in such a way that each data is plotted as point in n- dimensional space with the feature values represents the values of each co-ordinate.

E. Decision Tree

Decision tree is the learning model, which utilizes classifications problem. Decision tree module works by splitting the dataset into minimum of two sets. Decision tree’s internal nodes indicates a test on the features, branch depicts the result and leafs are decisions made after succeeding process on training.
Decision Tree works as follows

- Decision tree starts with all training instances linked with the root node
- It splits the dataset into train set and test set.
- It uses information to gain and chooses attributes to label each node. Subsets made contain information with a similar feature attribute.
- Above process is repeated till in all subset until leafs get generated in tree.

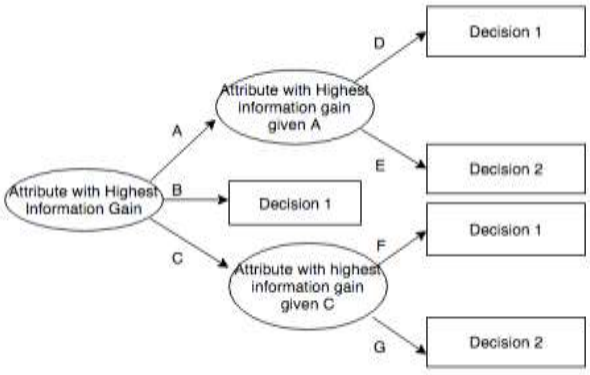

Figure 6: Decision Tree Model

The tree is constructed in such a way that no root to leaf node path contains same attribute twice. This is done repeatedly to construct every sub tree on the training instances, which is classified down through the path in the tree. For every record in the dataset, class label vaticinator problem starts with root of the tree. The root attributes are checked for the given record and then it checks next record attributes. This process continues till the value next node to go. The sample decision tree applied is depicted in Figure 6.
Implementation done as represented in the use case diagram given the figure 7.

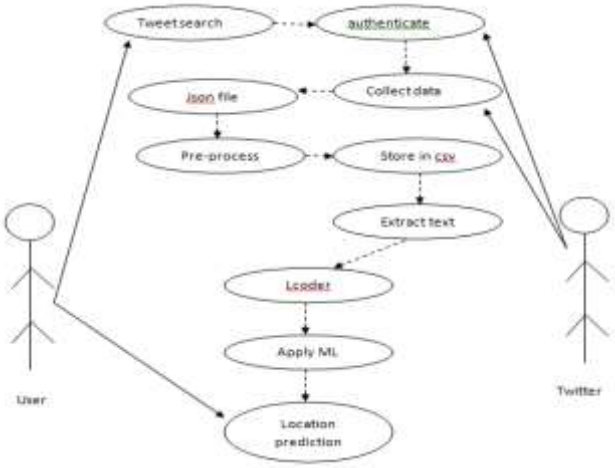

Figure 7: Decision Tree Model

The extracted features from the tweet are mentioned below code snippet.
(user["features"]["id"],user["features"]["name"],user["feat
ur
es"]["screen_name"],user["features"]["tweetstext"],user["f

ea
tures"]["HomePrecinct"],user["features"]["TweetPrecinct"
]
).
Instead of attaching the geo-tags to tweets, user may sometimes reveal the relevant precinct by specifying their name or landmarks in the tweets. During pre-processing the precinct names are important, thus we capitalize every words of tweet text to identify the geo-precincts. Geo precinct can be processed in two ways, one is through recognition, label the text and if recognized then they are converted to precinct. Next is through disambiguation, which makes the entries as identified precinct.

Table 1: Vaticinator Results

ID	Decision Tree	SVM	Naive Bayes
1	1	1	1
2	2	2	1
3	0	0	0
4	2	2	2
5	1	1	1
6	0	0	0
7	0	0	0
8	2	2	2
9	1	1	1
10	1	1	2

IV. RESULTS AND DISCUSSIONS

The pre-processed dataset is taken for machine learning process, we applied Naïve Bayes, S.V.M Pseudo-code and Decision Tree on the dataset (see table 1). The dataset is given 80% as training set and 20% as test set, we predicted the precinct and compared accuracy under following table 2.
The following table 2 shows the performance evaluation of three machine learning Pseudo-code namely Naive-Bayes, Support Vector machine (S.V.M) and Decision Tree. The evaluation parameters showed in the table are Accuracy of vaticinator. The table 2 clearly depicts that decision tree outperforms the other Pseudo-codes in terms of efficiency in accuracy.

Table 2: Accuracy Comparison

Pseudo-code	Accuracy
Naive-Bayes	43.67
S.V.M	86.78
Decision Tree	99.96

Table 3 shows the error rates in vaticinator. There are four error types calculated are Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE) and R-squared.

Table 3: Error Rate

Error Types	Naive-Bayes	S.V.M	Decision Tree
MAE	1.06	0.13	0.02
MSE	2.31	0.13	0.02
RMSE	1.52	0.36	0.04
R-Squared	0.01	0.88	1.00

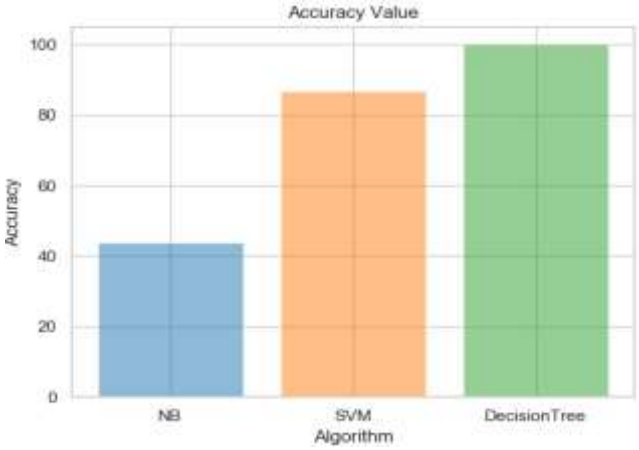

Figure 8: Performance Comparison

The above figur 8 shows the experimental results achieved using three machine learning Pseudo-codes. Naive-Bayes achieves around 40% of accuracy, S.V.M Pseudo-code achieves around 85% of accuracy and Decision Tree achieves around 99% accuracy. Thus from this work, we can conclude that Decision Tree is the suitable Pseudo-code for precinct vaticinator problem in tweet texts

V. CONCLUSION

Three precincts are considered from Social-Media data, namely home precinct, mentioned precinct and tweet precinct. When the Social-Media data is considered, geo-precinct vaticinator becomes a challenging problem. The tweet text nature and number of characters limitation make it hard to understand and analyze. In this work, we have predicted the geo-precinct of user from their tweet text using machine learning Pseudo-codes. We have implemented three Pseudo-codes to show the better performed one, which is suitable for geoprecinct vaticinator problem. Our experiment analysis concluded that decision tree is suitable for tweet text analysis and precinct vaticinator problem.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] Han, Bo & Cook, Paul & Baldwin, Timothy. (2012). Geoprecinct Vaticinator in Social Media Data by Finding Precinct Indicative Words. 24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers. 1045-1062.
[2] Ren K., Zhang S., Lin H. (2012) Where Are You Settling Down: Geo-locating Social-Media Users Based on Tweets and Social Networks. In: Hou Y., Nie JY., Sun L., Wang B., Zhang P. (eds) Information Retrieval Technology. AIRS 2012. Lecture Notes in Computer Science, vol 7675. Springer, Berlin, Heidelberg.
[3] Han, Bo & Cook, Paul & Baldwin, Timothy. (2014). Text-Based Social-Media User Geoprecinct Vaticinator. The Journal of Artificial Intelligence Research (JAIR). 49. 10.1613/jair.4200.
[4] Li, Rui & Wang, Shengjie & Chen-Chuan Chang, Kevin. (2012). Multiple Precinct Profiling for Users and Relationships from Social Network and Content. Proceedings of the VLDB Endowment. 5.

10.14778/2350229.2350273.
[5] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. 2014. Home Precinct Identification of Social-Media Users. ACM Trans. Intell. Syst. Technol. 5, 3, Article 47 (July 2014), 21 pages. DOI: http://dx.doi.org/10.1145/2528548
[6] Miura, Yasuhide, Motoki Taniguchi, Tomoki Taniguchi and Tomoko Ohkuma. "A Simple Scalable Neural Networks based Model for Geoprecinct Vaticinator in Social-Media." NUT@COLING (2016).
[7] A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, and M. M¨ uhlh¨ auser, "A multi-indicator approach for geolocalization of tweets," in Proc. 7th Int. Conf. on Weblogs and Social Media, 2013.
[8] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang, "Towards social user profiling: unified and discriminative influence model for inferring home precincts," in Proc. 18th ACM Int. Conf. on Knowledge Discovery and Data Mining, 2012, pp. 1023–1031.
[9] B. Han, P. Cook, and T. Baldwin, "A stacking-based approach to Social-Media user geoprecinct vaticinator," in Proc. 51st Annual Meeting of the Association for Computational Linguistics System Demonstrations, 2013, pp. 7–12.
[10] D. Flatow, M. Naaman, K. E. Xie, Y. Volkovich, and Y. Kanza, "On the accuracy of hyper-local geotagging of social media content," in Proc. 8th ACM Int. Conf. on Web Search and Data Mining, 2015, pp. 127–136.
[11] O. V. Laere, J. A. Quinn, S. Schockaert, and B. Dhoedt, "Spatially aware term selection for geotagging," IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 221–234, 2014.
[12] J. Mahmud, J. Nichols, and C. Drews, "Where is this tweet from? inferring home precincts of Social-Media users," in Proc. 6th Int. Conf. on Weblogs and Social-Media, 2012.