Journal Homepage: - www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI: 10.21474/IJAR01/4673
DOI URL: http://dx.doi.org/10.21474/IJAR01/4673
ISSN NO. 2320-5407

RESEARCH ARTICLE

CHARACTERISING USER DEMOGRAPHICS ACROSS SOCIAL NETWORK.

Charu Virmani[1] and Anuradha Pillai[2].
1. Research Scholar,YMCAUST, Faridabad.
2. Assistant professor, CE, YMCAUST, Faridabad.

Manuscript Info

Manuscript History

Received: 27 April 2017
Final Accepted: 30 May 2017
Published: June 2017

Key words:-
Social Media; Social Networks; Naïve Bayes Classifier.

Abstract

Analysis of the social media platform like Facebook & Twitter reveals a huge amount of information through user generated profile and comments. With the growing popularity of social media (Twitter), social network remains the largest as well as the most popular network. With registration of new users, tweets, news article, a user uploads their personal information or article and give his/her views about the some article. It contains huge amount of information about the user demographics, views about the video or article. This paper proposes a classification method for defining user behaviour by mining the user generated texts.

Introduction:-

From the early 2000s, user-generated content has become increasingly popular on the web; more and more users participate in content creation, rather than just consumption. Popular user- generated content (or social media) domains include blogs and web forums, social bookmarking sites, photo and video sharing communities, as well as social networking platforms such as Facebook and YouTube, which offers a combination of all of these with an emphasis on the relationships among the users of the community. Social media in general exhibit a rich variety of information sources in addition to the content itself; there is a wide array of non-content information available, such as links between items and explicit quality ratings from members of the community.

Much of the existing research on text information processing has been (almost exclusively) focused on mining and retrieval of factual information, e.g., information retrieval, Web search, and many other text mining and natural language processing tasks. Little work has been done on the processing of information about user using profile information until only recently. With the Web, especially with the explosive growth of the user generated content on the Web, the world has changed. One can generate information by registering on the network, post reviews of products at merchant sites and express views on almost anything in Internet forums, discussion groups, and blogs, which are collectively called the user generated content. Now if one wants to know the information about some user, it is no longer necessary to ask one's friends and families because there are plentiful of networks available where the information about the user can be extracted.

This paper comprises of an automated user demographics system, which extracts all the information from social network site on a particular generic and then make them into cluster according to user requirement preceded by data pre-processing like cleaning all the irrelevant data like URLs, misspellings and tagged names and then use these cluster for detecting user requirement. The paper is divided into 4 sections. Each section has a detailed description of the topic.

Corresponding Author:- Charu Virmani.
Address:- Associate professor, CSE, Manav Rachna International University, Faridabad.

Related Work:-

The online sites and the computerized impressions remain for eternity. By having solid connection between the client and their personality the online exceptional distinguishing proof of the client can be handled (Golder et al. 2014). Malhotra et al. (2012) exploits profile information for disambiguating user profiles across various social networks.

Tang et al. (2017) hosted two novel parameter tunable structures for beneficial spatial request taking care of with inadequate apportionments of Points of Interest (POI) on significant scale road frameworks. A uniform watchtower framework for beneficial question evaluation with a sensible stockpiling cost was proposed. Remembering the true objective to furthermore raise the question capability a hot-zone based watchtower framework by joining versatile customers advancement information was given into the physical structure of the improvement of watchtowers. Additionally a perfect watchtower association partition to achieve a desired congruity between the disengaged pre-figuring cost and the online question adequacy was gathered. Tests using this present reality road frameworks and geo social data demonstrated the power of the investigation system over the other bleeding edge approaches.

Mukhopadhyay et al. (2017) depicted the approach of looking in WWW were developing always yet the development rate of the changes were not that quick. The internet searcher proposed in this exploration ponders works viably and handles the difficulties of applicable not reachable site pages. In this review a model that uses various ontologies to play out different areas particular slithering for organizations to recognize their customers in the market were proposed. The proposed positioning calculation is another approach in light of the scaling component. This plan is versatile and could be effortlessly embraced by different undertakings as their device to perceive the users.

Sun et al. (2017) depicted the issue of securing the client protection in area sharing administrations, for example, adjacent companions' inquiry and outsider's question. Another system and another inquiry calculation (UDPLS) were proposed to ensure client area protection on informal organization server and client's interpersonal organization security on area protection. The client can impart the area to determined companions rather than all companions. The question time of the structure nearly has no impact on the quantity of companions in the companion inquiry. The nom de plumes client's companions and ID in the client terminals were coordinated. The broad recreation tests assessed the execution of the proposed calculation. The proposed examine work brought about extra movement overhead. None of the author has taken advantage of profile information to characterize the user's information on the basis of user's demographics

Proposed Work:-

Naive bayes prediction classifier is a novel classification approach for analysis of new user's demographics posted on the social media websites. Profile information is an informative way to exchange information, views and connect with the audience in an interactive way.

Architecture of proposed System:-

Architecture of proposed system is shown in Figure 3.1.

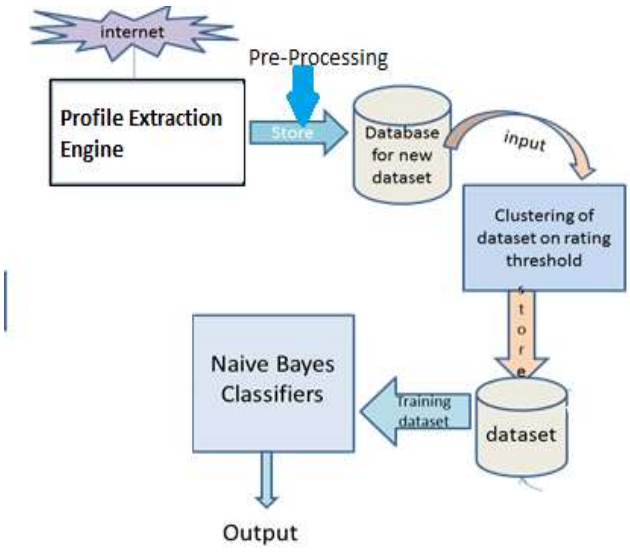

Figure 3.1:- Architecture of proposed system

The proposed system consists of four components. These components are as follows:

Profile Extraction Engine:-

This engine takes query "q" as a input and it will extract all the information from all the profiles on that term q and it will automatically store all these information with all its Meta data information into a database. All these information like user's Name, age, Location, groups, interest and connections, posts and information regarding who posted that comment on that particular video or article (like in case of yahoo news). For the purpose of research twitter is taken as input sites.

Data Pre-processing:-

In this phase of the proposed system all uncleared message like URL and other unknown words from the profile information are removed. Spelling checking and removal of all the stop words from both clusters set are also done here.

Clustering of users as per user's need:-

In this phase of the proposed system all the dataset is taken as an input and clustering is done on the profiles as per user requirement. Clustering of dataset is done using user's interest threshold here.

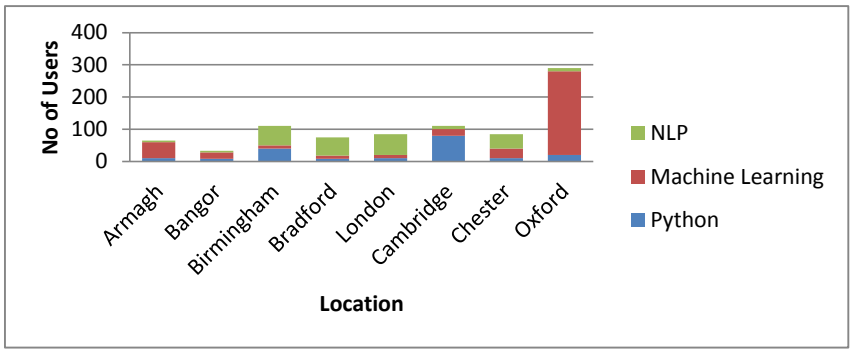

Figure 2. Count of user for different skills for different location

Two partition sets have been taken here. In one set all the profiles with post or profile information are kept which we consider as profile acceptable as per user's query and in other set we will keep all profiles that are not related to the query.

Train the naïve bayes prediction Classifier:-

This classifier takes both the clusters and train naive bayes classifier. After training the naive bayes classifier any new input profile is input to this system to detect all if the user is acceptable.

Algorithm for proposed system:-

1. let P be our search term or generic on which we collect profile set D_i using extraction engine or manually .
2. Q_p and Q_n are the threshold values that used for clustering into profile set ACT_i & REJ_i .
3. do pre-processing on ACT & REJ sets for spelling and stop word removal.
4. for $\forall i$ in set D
   If $D_i \geq Q_p$
   i. Put D_i in ACT_i
   If $D_i \leq Q_n$
   ii. Put D_i in REJ_i
5. make a array as Pre[T_i][Setname].
6. for $\forall C_i$ in ACT set
   Calculate Pre[T_i][ACT].
7. for $\forall C_i$ in REJ set
   Calculate Pre[T_i][REJ].
8. let new profile CT for which we want to check acceptability apply naive bayes theorm
   On each term T_i .

$$P(x_j|d) = \frac{\left(\Pi\, P(t_i|x_j)\right) * P(x_j)}{P(d)}$$

Here x_j are the set of user's information.

The proposed algorithm works on the naive bayes theorem. In this algorithm p is a topic or generic for which we want to collect our dataset. After collecting the dataset D from profile extraction engine, data preprocessing is applied on these two sets to remove garbage text. It then starts training our prediction classifier by creating an array Pre[T_i][Setname].we do clustering on this dataset and divide it on the basis of two threshold values of acceptable and unacceptable count Q_p and Q_n into two sets ACT_i & REJ_i. After training the classifier by both sets ACT_i & REJ_i the prediction classifier is used for new user to predict the acceptability to community.

Conclusion and Future Scope:-

In this research work, in-depth analysis of twitter users has been conducted to shed some light on different aspects of user demographics. Large-scale studies using twitter meta data revealed strong dependencies between different kinds of user's interest, user information provided by the networks and topic orientation of the discussed content. The proposed techniques have direct applications to user specific search. In future, this proposed system can be automated using Google API to that every user can get benefit of this technique. We plan to study additional stylistic and linguistic features, relationships between users and techniques for aggregating information obtained from multiple social networks. This could for instance be applied for identifying groups of users with similar interest and recommending contacts or groups to users in the system.

References:-

1. Mukhopadhyay, D., & Kulkarni, S. (2017). An Approach to Design an IoT Service for Business—Domain Specific Web Search. In *Proceedings of the International Conference on Data Engineering and Communication Technology* (pp. 621-628). Springer Singapore.
2. Sun, G., Xie, Y., Liao, D., Yu, H., & Chang, V. (2017). User-defined privacy location-sharing system in mobile online social networks. *Journal of Network and Computer Applications*, *86*, 34-45. [3] F. Wu and B. A.
3. Tang, L., Chen, H., Ku, W. S., & Sun, M. T. (2017). Exploiting location-aware social networks for efficient spatial query processing. *GeoInformatica*, *21*(1), 33-55.
4. Malhotra, A., Totti, L., Meira Jr, W., Kumaraguru, P., & Almeida, V. (2012, August). Studying user footprints in different online social networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on* (pp. 1065-1070). IEEE.
5. Golder, S. A., & Macy, M. W. (2014). Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology*, *40*, 129-152.
6. S. Chakrabarti. Mining the Web: Discovering Knowledge from Hypertext Data. Morgan-Kauffman, 2002.
7. S.M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. Automatically assessing review helpfulness. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 423–430, Sydney, Australia, July 2006. Association for Computational Linguistics.
8. Gull, Angadi, dr.santoshkumar gandhi : tracing high quality content in social media for modelling & predicting the flow of information – a case study on facebook. In: international journal of emerging trends & technology in computer science volume 2 issue 2, April 2013
9. Zhou, M., Zhang, W., Smith, B., Varga, E., Farias, M., & Badenes, H. (2012, February). Finding someone in my social directory whom i do not fully remember or barely know. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces* (pp. 203-206). ACM.
10. Noor, S., & Martinez, K. (2009, June). Using social data as context for making recommendations: an ontology based approach. In *Proceedings of the 1st Workshop on Context, Information and Ontologies* (p. 7). ACM.
11. Mendes, P. N., Jakob, M., & Bizer, C. (2012, May). DBpedia: A Multilingual Cross-domain Knowledge Base. In *LREC* (pp. 1813-1817).
12. Dalvi, B., Minkov, E., Talukdar, P. P., & Cohen, W. W. (2015, February). Automatic gloss finding for a knowledge base using ontological constraints. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (pp. *369-378). AC*