High-value consumer profiling using a hybrid text-based and web-based approach

Kalyani Jagdale
Georgia Institute of Technology
kalyanijagdale@gatech.edu

Rishi Bubna
Georgia Institute of Technology
rishi.bubna@gatech.edu

Uma Sreeram
Georgia Institute of Technology
uma.sreeram@gatech.edu

ABSTRACT

Social media websites like Twitter and Facebook hold humongous data of their users and present an opportunity for businesses to convert this data into valuable insights about their customer profiles. However, it is a challenging task to analyze large volumes of social media activity to identify high-value potential users that are likely to be interested in a specific product or service.

In this project, we propose an approach to find out new potential consumers for a specific product or service offering using social media presence of consumers. Using a hybrid, text-based and web-based approach we aim to identify and rank high-value customers using text-analysis, network-analysis and machine learning.

This allows businesses to better design customer engagement programs targeting the right social media audience that are most likely to convert into consumers, thus improving the efficiency of customer acquisition, and increasing return on investment.

KEYWORDS

consumer profiling, text-analysis, web-based analysis, high-value users, social media

1 INTRODUCTION

This project focuses on identification of high-value user profiles on social media platforms such as Twitter for a particular business. A social media user is identified as a potential high-value consumer, if there is evidence that the user is interested in the content shared by the business social media account.

The hypothesis here is that followers of a particular business social media account are interested in the content posted and the products/services the business has to offer - hence they choose and take action to follow the business account.

Therefore, in that case, as the interests of the followers are similar to that of the business account, the content shared by the follower is also expected to be of a similar nature to the content shared by the business account.

In other words, the content shared by the business social media account can be used to identify the group of followers who are interested in the content that the business has been posting. Thus, this project uses text-based analysis to identify similarity between content posted by the business account, and other social media users to find the right target audience that are likely to be interested in the offerings by a business compared to others who are not sharing similar contents, and may not be interested in the business offerings.

However, not every social media user is highly active, or posts considerable amount of content such that an effective text-based analysis can be performed to analyze their behavior. The figure below depicts the number of tweets for 1000 randomly selected Twitter users. [1]

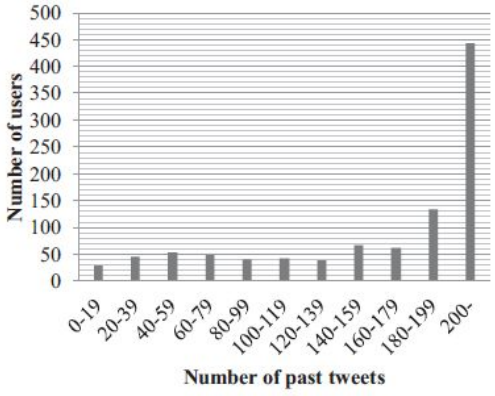

Figure 1: Distribution of number of Tweets of users

The figure shows that more than half of the twitter users have made less than 200 tweets. Thus, substantiating that text-based analysis of social media content may not be sufficient to profile the majority of social media users.

Therefore, in this project we propose a hybrid approach that combines text-analysis with web-based network analysis to

profile potential consumers even with low social media posts.

Every business that has an online presence and that maintains social media accounts can benefit from this project. This will allow businesses to identify potential high-value consumers that are most likely to respond to their customer-engagement programs, leading to high efficiencies in customer acquisition.

2 LITERATURE REVIEW

Previous work in customer profiling using social media platforms such as Twitter is discussed in the research by Yibin Yang, 2019 [2]. The authors use a text-based approach to find out similar tweets by (business) account owners and its followers. They use a mining pipeline that consists of preprocessing, feature selection and data resampling. Using Twitter LDA, the authors perform topic modelling to reduce the dimensionality of tweets, followed by the use of support-vector machine technique to identify high-value consumers.

Similarly, in previous works of Siao Ling Lo in 2015 [3] and 2016 [4], the authors proposed an approach to identify high-value audiences on Twitter without having to manually annotate the vast amount of tweet contents. They deployed a Fuzzy keyword matching technique to generate the similarity score that results in an improved performance, even compared to the work done later by Yibin Yang, 2019 [2].

However, the approaches discussed above are only limited to content information which are tweets posted publicly by users on their Twitter accounts. Further, these techniques do not effectively capture the behavior of users who post insignificant online content.

A hybrid text-based and community-based approach was presented in the previous work by Kazushi Ikeda in 2013 [1]. The authors focus on estimating the demographics of users using a hybrid of text-based analysis and community-based analysis to identify the characteristics of both types of users who tweet frequently, and infrequently. For a target user who posts infrequently, a user community of high frequency tweeters is extracted from the followers and followees of the target user and are clustered into groups. Then the demographic category of each community group is estimated by analysing the demographic distribution of the group member, to estimate the demographic of the infrequent target user.

3 PLAN OF ACTION

For the purpose of this project we plan to analyze social media behaviour of users on Twitter. We will use Twitter API to obtain information about a user's social profile (their followers and friends) and their previous tweets. As also presented in the works of Yibin Yang 2019 [2], we aim to create a representative set of both the target account (which represents our focus business account) and a non-target account (with accounts of different categories) to accurately find similarity between posts of users and business accounts to classify and identify high-value consumers.

As discussed in [3], we will use text mining approaches like Twitter LDA and SVM to calculate the similarity score in tweets. Further, as previously presented by Kazushi Ikeda, 2013 [1], we plan to use a hybrid approach where text-based approach will be used for users who tweet frequently and community network-based technique will be used to identify the characteristics of the users who tweet infrequently. For the pool of non-active users, user communities of high frequency tweeters can be extracted from the followers and clustered into groups. Then the value of each community group can be estimated by analysing the value distribution of the group members.

Using model performance metrics like precision, recall we will assess the model formulations and appropriately tune the parameters and techniques of our approach to further improve the performance of our model.

However, there are possible challenges that exist in the approach of this project, such as inaccurate information encoding and possible unbalanced information being used to train the classifier. We will attempt to implement various vector representations to accurately encode information of the results from the topic modelling phase, in order to accurately classify high value users. Also, we will attempt different sampling techniques to deal with unbalanced data to ensure that the classifier is unbiased.

REFERENCES

[1] Ikeda, K., Hattori, G., Ono, C., Asoh, H. and Higashino, T., 2013. Twitter user profiling based on text and community mining for market analysis. Knowledge-Based Systems, 51, pp.35-47.

[2] Yang, Yibing, and M. Omair Shafiq. "Identifying High Value Users in Twitter Based on Text Mining Approaches." In 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), pp. 634-641. IEEE, 2019.

[3] Lo, S.L., Cornforth, D. and Chiong, R., 2015. Identifying the high-value social audience from Twitter through text-mining methods. In Proceedings of the 18th Asia Pacific Symposium on Intelligent and Evolutionary Systems, Volume 1 (pp. 325-339). Springer, Cham.

[4] Lo, S.L., Chiong, R. and Cornforth, D., 2016. Ranking of high-value social audiences on Twitter. Decision Support Systems, 85, pp.34-48.
[5] Okazaki, S., Díaz-Martín, A.M., Rozano, M. and Menéndez-Benito, H.D., 2015. Using Twitter to engage with customers: a data mining approach. Internet Research.
[6] Raghuram, M.A., Akshay, K. and Chandrasekaran, K., 2016. Efficient user profiling in twitter social network using traditional classifiers. In Intelligent systems technologies and applications (pp. 399-411). Springer, Cham.