IDENTIFICATION OF TWITTER COMMUNITIES INTERESTED IN AFRICAN AFFAIRS

INTRODUCTION

Twitter has recently emerged as a popular social microblogging service where users share and discuss about everything, including news, jokes, what they are going through, and even their mood. Nowadays there are over 100 million users in Twitter, and almost 200 million messages are posted every day.

The importance of this research include but are not limited to the following

- Tweet or User recommendation: prominent individuals in a field can be identified and subsequently suggested to other users interested in that field
- Targeted Marketing: this form of marketing is directed towards some group of people that can be identified via their twitter community
- Sponsorship and Ambassadorial position: influencers in a particular field can be identified for ambassadorial position

METHODOLOGY

Data sources and description

In this study, twitter API serves as the main link to information. However Google API provided the first set of users that are prominent in field such as Journalism, Entertainment and Government outfits. Twitter API was then used to obtain more user (friends) from the obtain handles. The timeline of this handles were mined and used in subsequent analysis.

With the purpose of computing user similarity, we leverage different features which reflect one's interests, including both textual contents and social structure. The textual contents in Twitter mainly encompass three features: tweet text, URLs and hashtags embedded in tweets. As we know, tweet text contains rich information about author of the tweet, such as what the user is talking about, what the user is going through, etc. So tweet text is potentially useful in determining interests of an individual user. The mined tweets contained few URLs and hashtags, as such the tweet text serve as the main source of information.

COMMUNITY IDENTIFICATION

The goal of this section is automatically identifying the topics that users are interested in based on the tweets they published and computing tweet text similarity based on these topics. To avoid the problem of small size of a single tweet and the sparseness of data, we aggregate the tweets published were compressed into a document. Thus, the document essentially contains tweets by users, and finding topics that users are interested in just means finding latent topics in this document. With this purpose, Latent Dirichlet Allocation (LDA) model is applied, which is an unsupervised machine learning method to identify latent topics from large document collections.

As a generative probabilistic model, LDA generates each document as follows: first, pick a topic from its distribution over topics for each document; second, sample a word from the chosen topic's distribution; finally, repeat the two above processes for all the words in a document. Thus, each document is represented as a probability distribution over some topics, while each topic is represented as a probability distribution over a bag of words. The keywords generated were used in labelling the tweets

Clustering algorithm (kMeans) was used to identify the cluster of communities. A minor difficulty in applying k-means is the selection of parameter k, which decides the number of clusters and has a great impact on the final results of community discovery. This will be discussed subsequently.

RESULTS

About 500 tweet details from 1000 users with more than 500 followers were collected and used in the analysis.

The LDA algorithm was able to produce keywords that enable the classification of the tweets into the following categories **economy, social, cultural, health.** The users were aggregated based on the number of their corresponding tweets in each category such that a user with 200 economy tweets, 150 health tweets, 150 social tweets, and 100 cultural tweets will have a coordinate value of (200, 150, 100, 150).

In kMeans clustering, the optimal k was selected using silhouette score. The k with the highest score (k=2) was selected as optimal number of clusters (Figure 1)

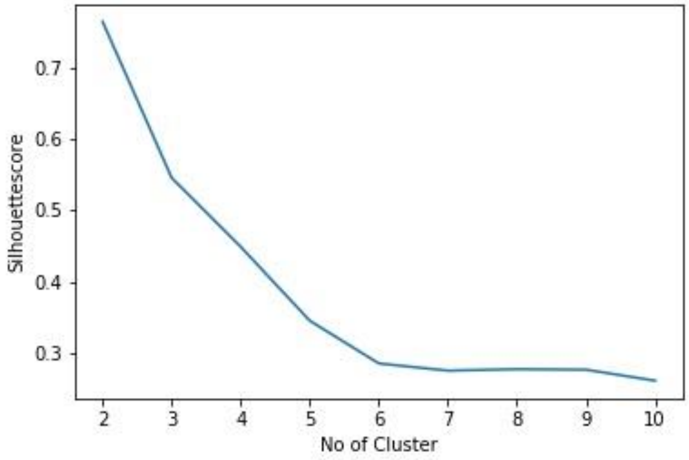

Figure 1: A diagram showing Silhouette score at different k values

kMeans clustered the data into two distinct groups. These groups are overlaps of communities. The cluster are interpreted as relationships between the communities (Figure 2). The cluster reveals a difference between active contributors to the above topics and user with fewer tweets. The first group (active contributors) constitute about 81.9% while the other group is about 18.1%. The majority of users in the first groups are government outfits, government officials, entertainers and online influencers. The second group contain users who are not quite interested in the identified topics but contribute once in a while to trending topics.

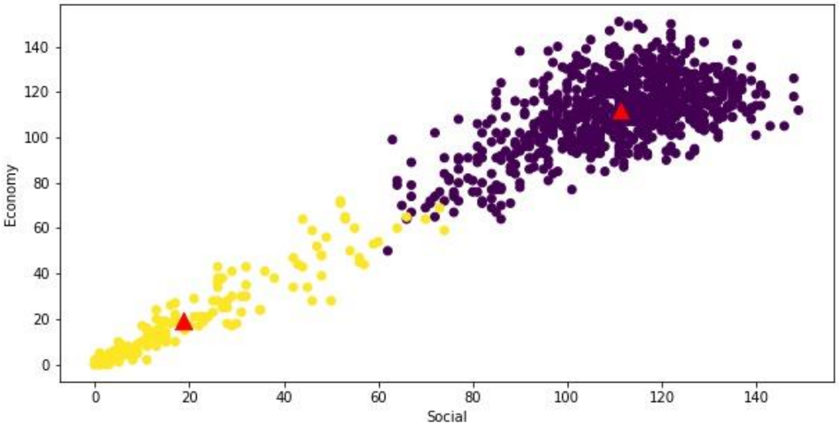

Figure 2: A plot showing cluster relationships

NB: The red markers denotes the center of the clusters

It must be noted that about 5% of tweets obtained contained hashtags related COVID 19. Less than 1% of the first groups used hashtags that are related to COVID 19. A recognizable amount of the hashtags was related to society issues and Government or antigovernment policy related events. About 0.08% of the hashtags were related to holidays. Some of the tweets carried hashtags that included ‘Nigeria’ in them.

CONCLUSION

A fundamental task of understanding Twitter at user level is community discovery. It can be inferred that there is a relationship between twitter communities and tweets can be used to obtain this relationship among the communities.