12-2014

Identifying the high-value social audience from Twitter through text-mining methods

Siaw Ling LO
Singapore Management University, sllo@smu.edu.sg

David CORNFORTH

Raymond CHIONG

Identifying the High-Value Social Audience from Twitter through Text-Mining Methods

Siaw Ling Lo[1,2], David Cornforth[1] and Raymond Chiong[1]

[1]School of Design, Communication and Information Technology, The University of Newcastle, Callaghan, NSW 2308, Australia
[2]School of Information Technology, Nanyang Polytechnic, Singapore

siawling.lo@uon.edu.au
{david.cornforth,raymond.chiong}@newcastle.edu.au

Abstract. Doing businesses on social media has become a common practice for many companies these days. While the contents shared on Twitter and Facebook offer plenty of opportunities to uncover business insights, it remains a challenge to sift through the huge amount of social media data and identify the potential social audience who are highly likely to be interested in a particular company. In this paper, we analyse the Twitter content of an account owner and its list of followers through various text mining methods, which include fuzzy keyword matching, statistical topic modelling and machine learning approaches. In order to reflect the real-world scenario, the tweets of the account owner are used to segment the list of the followers and identify a group of high-value social audience members. This enables the account owner to spend resources more effectively by sending the right offers to the right audience and hence maximize marketing efficiency and improve the return of investment.

Keywords: Twitter, Topic modelling, Machine learning, Audience segmentation

1 Introduction

Social media has not only transformed the way we share our personal life, it has also transformed the way business is carried out. A recent study [1] found that nearly 80% of consumers would more likely be interested in a company due to its brand's presence on social media. It is therefore not a surprise that 77% of the Fortune 500 companies have active Twitter accounts and 70% of them maintain an active Facebook account to engage with their potential customers [2]. With more companies doing businesses on social media, how can one stand out from the increasingly crowded social space to find prospective customers from the audience in social media?

It is no longer feasible for a company to depend on gimmicks (such as incentive referrals) to boost the social media business as that may only provide short-term gain. While a company can adopt approaches like mass marketing to all the "fans" or con-

tacts available, the return may not be justified by the effort and amount of money spent. Furthermore, there is a thin line between broadcasting a general message and spamming, so instead of attracting a greater audience, there is a high probability of losing current customers. Hence, it makes sense to identify a target audience to maximize the marketing efficiency and improve the return of investment (ROI).

Traditionally, an understanding of customers is obtained through customer surveys so that information such as customer preferences can be known. This set of information can be merged with internal company data, for example, product purchase data or transactional data, so that segmentation of customers can be done to better understand the customers and manage offerings according to their interests. However, with the recent proliferation of social media activities, more and more companies are putting in efforts to ensure that their presence is felt in the crowded social space. Even though there is a rich source of customer information to be mined, the real-time nature and free-form expression poses a challenge in extracting commercially viable contents from the vast amount of conversations.

In order to reach out to potential customers, companies use mailers or emails to inform them about their new products or promotions from third party or internal listings. With the proliferation of social media, companies are now using Facebook fan pages or Twitter accounts to engage with their fans and followers or social audience. There are currently two methods in identifying or reaching to the audience on social media – keyword search and semantic tagging.

While there are many guides or tips on the web on how to find the target audience on social media, most of these concentrate on searching specific keywords related to products or brands. However, while using this approach can retrieve lists of information using different keywords, it is not capable of determining the relationship among the keywords and providing a more comprehensive view on the subject matter without the help of domain experts. Furthermore, deciding which keywords to use may not be obvious to a non-expert and this may lead to inaccurate information extraction and hence a misunderstood market analysis. On top of this, there is a need to manually consolidate the list of social audience found and to ensure that the content shared by the audience matched with the keywords.

Prior work [3][4] has proposed various approaches such as translating both social networks and semantic information into Resource Description Framework (RDF) formats and using RDF methods for correlation, or the use of semantic tagging to correlate the current social tagging approach to make sense of the social media data. These approaches, however, require additional efforts of translating and tagging of current social media data, which can be a daunting task considering the huge amount of data and the possible manual effort.

In this paper, we investigate several different methods in order to make use of available resources to identify a group of high-value social audience members without utilizing a considerable amount of human annotation effort. These include text mining methods such as fuzzy keyword matching using Dice coefficient [5] of string similarity, statistical topic modelling with Twitter Latent Dirichlet Allocation (LDA) [6], and machine learning using the Support Vector Machine (SVM) [7]. The hypothesis is based on the idea that the followers are interested in the content posted and hence they

choose and take action to follow the account owner. If that is the case, some of the tweets shared by the followers should be of similar nature to the account owner. In other words, the tweets of the account owner (of a similar period of time) can be used to select or identify the group of followers who are interested in the content that the owner has been tweeting. Hence, these followers are more likely to comprise the target audience compared to others who are not sharing similar contents.

In order to achieve this, we use a list of seed words (derived from the owner tweets using term frequency analysis) to generate a baseline using Direct keyword matching. This set of seed words is used in Fuzzy matching and identification of suitable topic numbers from Twitter LDA. In contrast to a traditional machine learning approach, tweets from the account owner are used to build the positive training data instead of the tweets extracted from the list of followers. This eliminates the need to manually annotate the vast amount of tweets from the followers and it is more practical if it is to be adopted in a real-world application.

The major contributions of this work are as follows:

- To the best of our knowledge, our work in this paper is the first attempt to identify the target audience from the list of followers of a Twitter owner's tweets through various methods. It is assumed that those who tweeted similar contents are more likely to be interested in the owner's tweets, compared to others who have not been sharing similar contents.
- From the result observation, it is likely that half or less followers are tweeting similar contents as the owner. This implies that it may not be sensible to try to engage every follower, as not everyone is interested in the content or topic shared. Instead, it makes sense to be selective and target specific groups of followers to maximize the use of allocated marketing expenses and reach out to potential customers in social media.

2 Related Work

As the aims of any business are to increase profit, build a long lasting brand name, and to grow the customer base or engage current customers, it is essential to understand the needs and behavior of the customers. This understanding can be achieved through different means and at different levels of detail. Most companies define a set of segments that reflect the companies' knowledge of the customers and their traits or behavior. All other marketing activities, such as customer engagement activities, are targeted and measured according to this segmentation.

However, this segmentation is typically restricted to customer relationship management (CRM) or transaction data obtained either through customer surveys or tracking of product purchases to understand the customer demand. Demographic variables, RFM (recency, frequency, monetary) and LTV (lifetime value) are the most common input variables used in the literature for customer segmentation and clustering [8, 9]. While these internal systems can be coupled with geographical data to ob-

tain additional information, it remains limited to the specific system and does not leverage on the sharing and activities on social media where customers tend to reveal about themselves – life events, personal and business preferences, perception of brands and more.

There are efforts in deriving or estimating the demographics information [10, 11] from the available social media data, but this set of information may not be able to be used directly in targeted marketing, as temporal effect and type of products to be targeted are usually not considered. Besides that, the demographic attributes such as age, gender and residence areas may not be updated and hence may result in a misled conclusion. Recently, eBay has expressed that, due to the viral campaigns and major social media activities, marketing and advertising strategies are evolving. Targeting specific demographics through segmentation, although still have its value, eBay is focusing on "connecting people with the things they need and love, whoever they are" [12]. Other research on predicting purchase behavior from social media has shown that Facebook categories, likes and n-grams significantly outperform demographic features shared on Facebook [13]. Due to the privacy policy of Facebook profiles, this paper focuses on Twitter, where most of the content and activities shared online are open and available. It is interesting to identify if other factors (such as the text content shared on social media) can be used to derive alternative approaches to identify the target or high-value social audience for a company or a product.

3 Methods

The focus of this research is to establish an approach that makes use of contents and activities shared on social media platforms to profile and segment the social audience of a Twitter account owner. This account owner can be a business or a government body and the online social audience we are interested to profile or segment is the list of followers of the Twitter account. The architecture of our system is given in Fig. 1.

The tweets from various parties - owners', followers', owners' from other domains are cleaned and preprocessed before preparing for seed words generation and SVM training and testing datasets. The owners' tweets are used as the positive training data while tweets of owners' from other domains are extracted as the negative training data. 10 fold cross-validation is applied on both the positive and negative training datasets for the SVM model before the classification of followers' tweets (or the testing data) is conducted. The seed words generated are used in both Fuzzy Keyword Match and Twitter LDA methods. A string similarity score derived from Dice coefficient is calculated through a fuzzy comparison with the seed words on the testing data. A list of topics is learnt from testing data using the Twitter LDA and followers with relevant topic numbers are identified. Details of each component are described in the following sections.

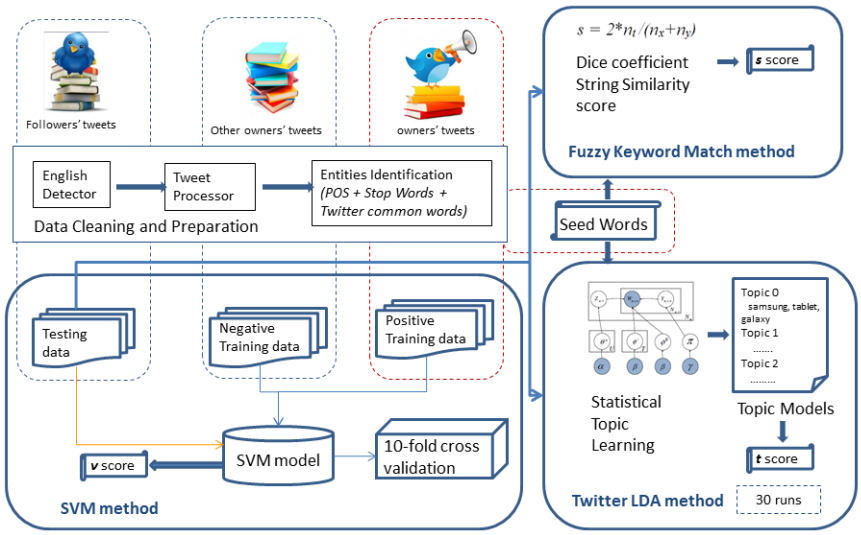

Fig. 1. System architecture.

3.1 Data Collection

We use the Twitter Search API [14] for our data collection. As the API is constantly evolving with different rate limiting settings, our data gathering is done through a scheduled program that requests a set of data for a given query. The subject or brand selected for this research is Samsung Singapore or "samsungsg" (its Twitter username). At the time of this work, there were 3,727 samsungsg followers. In order to analyze the contents or tweets of the account owner, the last 200 tweets by samsungsg have been extracted. The time of tweets ranges from 2 Nov 2012 to 3 Apr 2013. For each of the followers, the API is used to extract their tweets, giving a total of 187,746 records, and 2,449 unique users having at least 5 tweets in their past 100 tweets of the same period. We reasoned that those with fewer than 5 tweets were inactive in Twitter, as it implied that the user was tweeting an average of less than one tweet in a month (since the period was of 6 months).

3.2 Data Cleaning and Preparation

Tweets are known to be noisy and often mixed with linguistic variations. It is hence very important to clean up the tweet content prior to any content extraction:

- Non-English tweets are removed using the Language Detection Library for Java [15];
- URL, any Twitter's username found in the content (which is in the format of @username) and hashtags (with the # symbol) are removed;
- Each tweet is pre-processed to lower case.

As tweets are usually informal and short (up to 140 characters), abbreviation and misspelling are often part of the content and hence the readily available Named Entity Recognition (NER) package may not be able to extract relevant entities properly. As such, we derive an approach called Entities Identification, which uses Part-of-Speech (POS) [16] tags to differentiate the type of words. All the single nouns are identified as possible entities. If the tag of the first fragment detected is 'N' or 'J' and the consecutive word(s) is of the 'N' type, the word(s) will be extracted as phrases. This approach is then complemented by another process using the comprehensive stop words list used by search engines (http://www.webconfs.com/stop-words.php) in addition to a list of English's common words (preposition, conjunction, determiners) as well as Twitter's common words (such as "rt", "retweet" etc.) to identify any possible entity. In short, the original tweet is sliced into various fragments by using POS tags, stop words, common words and punctuation as separators or delimiters. For example, if the content is "Samsung is holding a galaxy contest!", two fragments will be generated for the content as follows: (samsung) | (galaxy contest).

3.3 Seed Words Generation

All the tweets extracted from samsungsg are subjected to data cleaning and preparation mentioned in the previous section. Each tweet is now represented by the identified fragments or words and phrases. This set of data is further processed using term frequency analysis to obtain a list of seed words (which include "samsung", "galaxy s iii", "galaxy camera" etc.). The words in a phrase are joined by '_' so that they can be identified as a single term but the '_' is filtered in all the matching processes.

These seed words are used to generate results for Direct Keyword Match, Fuzzy Keyword Match and identification of suitable topic numbers in the Twitter LDA method (see Section 3.6).

3.4 Direct Keyword Match

This is the most common method used to find the relevant or suitable social audience for a specific content or product. The list of seed words generated is used to match the tweets from the list of followers. As long as there is a direct word or phrase match with any of the seed words, the follower will be considered as a potential member of a high-value social audience, who is likely to be interested in the content shared by the account owner. The result of this approach is set as the baseline for the rest of the methods.

3.5 Fuzzy Keyword Match

It is not uncommon for Twitter users to use abbreviations or interjections or a different form to represent a similar term. For example, "galaxy s iii" can be represented by "galaxy s 3", which is understandable by a human but cannot be captured by the Di-

rect Keyword Match baseline method. As such, a Fuzzy Keyword Match method using the seed words derived is implemented.

The comparison here is based on a Dice coefficient string similarity score [5] using the following expression,

$$s = 2*n_t/(n_x+n_y) \quad (1)$$

where n_t is the number of characters found in both strings, n_x is the number of characters in string x and n_y is the number of characters in string y. For example, to calculate the similarity between "process" and "proceed":

x = process	bigram for x = {pr ro oc ce es ss}
y = proceed	bigram for y = {pr ro oc ce ee ed}

Both x and y have 6 bigrams each, of which 4 of them are the same. Hence, the Dice coefficient string similarity score is 2*4/(6+6) = 0.67.

Similar to the Direct Keyword Match method, each of the tweets of every follower is compared with the seed words and the highest score of any match is maintained as the *s* score of the follower.

3.6 Twitter Latent Dirichlet Allocation (LDA)

Recently, Latent Dirichlet Allocation (LDA) [17], a renowned generative probabilistic model for topic discovery, has been used in various social media studies [6][18]. LDA uses an iterative process to build and refine a probabilistic model of documents, each containing a mixture of topics. However, standard LDA may not work well with Twitter as tweets are typically very short. If one aggregates all the tweets of a follower to increase the size of the documents, this may diminish the fact that each tweet is usually about a single topic. As such, we have adopted the implementation of Twitter LDA [6] for unsupervised topic discovery among all the followers.

As the volume of the tweet set from all the followers is within 200,000, we have chosen to learn a smaller number of topics (from 10-50) from Twitter LDA. These 5 different topic models run for 100 iterations of Gibbs sampling while keeping the model parameters or the Dirichlet priors to be the same, where α: 0.5; β_{word}: 0.01, $\beta_{background}$: 0.01 and γ:20. The suitable topics are chosen automatically via comparison with the list of seed words. The result or the list of audience identified by each topic model is a consolidation of 30 runs where a score is assigned to each follower using the following calculation:

$$t = n_m/n_r \quad (2)$$

where n_m is the total number of matches and n_r is the total number of runs. If a particular follower is found in 5 runs then the *t* score assigned is 5/30 = 0.17.

3.7 Support Vector Machine (SVM)

The SVM is a supervised learning approach for two- or multi-class classification, and has been used successfully in text categorization [7]. The SVM separates a given known set of {+1, -1} labeled training data via a hyperplane that is maximally distant from the positive and negative samples respectively. This optimally separating hyperplane in the feature space corresponds to a nonlinear decision boundary in the input space. More details of the SVM can be found in the literature [19].

The positive dataset is generated using processed tweets from the account owner or samsungsg. The negative dataset is randomly generated from account owners of 10 different domains (online shopping deals, food, celebrities, parents, education, music, shopping, politics, Singapore news, traffic), which are ilovedealssg, hungrygowhere, joannepeh, kiasuparents, MOEsg, mtvasia, tiongbahruplaza, tocsg (TheOnlineCitizen), SGnews and sgdrivers respectively. These domains are chosen as they are the main topics discovered from Twitter LDA from the list of tweets of all the followers. The respective account owners are selected as they are the popular Twitter accounts in Singapore according to online Twitter analytic tools such as wefollow.com.

LibSVM implementation of RapidMiner [20] is used in this study and the sigmoid kernel type is selected as it produces higher precision prediction than other kernels, such as RBF and polynomial.

As the number of tweets shared by each follower is different, there are various approaches in representing followers' tweets as the testing data:

- Extract topical representation features of all the tweets from each follower using the top topical words from Twitter LDA;
- Extract word representation features of all the tweets from each follower using term frequency;
- Treat each set of followers' tweets as individual testing data, where each tweet will be classified as either positive or negative. The final assignment of the *v* score is based on the following representation:

$$v = n_p/n_a \tag{3}$$

where n_p is the total number of tweets that are classified as positive and n_a is the average number of tweets shared by all the followers (71 tweets per follower for this study). If 5 tweets of a particular follower is classified as positive, then the *v* score assigned is 5/71 * *normalized factor* so that the score range is within [1, 0].

4 Experiments and Results

The results obtained from the various methods were compared with a random annotated sample of the followers. The contents of a total of 300 followers (which were randomly sampled) were annotated manually as either a potential high-value social audience according to the content shared by the account owner or not a target audience. This set of data was used in the evaluation of the various methods described in Sections 3.4, 3.5, 3.6 and 3.7.

4.1 Numbers of High-Value Audience Identified

Out of the 2,449 'active' followers (excluding those tweeted less than 5 tweets), the numbers of the followers who were tweeting similar contents measured by various methods are listed in Table 1.

Table 1. Numbers of high-value audience identified by various methods

Methods	Audience Numbers	% within active followers (2449)	% within all the followers (3727)
Direct Keyword Match	321	13%	9%
Fuzzy Keyword Match	1115	46%	30%
Twitter LDA 10 topics*	760	31%	20%
Twitter LDA 20 topics*	582	24%	16%
Twitter LDA 30 topics*	527	22%	14%
Twitter LDA 40 topics*	414	17%	11%
Twitter LDA 50 topics*	424	17%	11%
SVM	736	30%	20%

*The results are consolidated from 30 runs.

4.2 Results of Twitter LDA

As shown in Table 1, in general, the size of high-value audience decreased with the increase of the topic numbers. A further analysis was done and the group of audience members identified with topic numbers greater than 30 remained the same and hence further result analysis is done on topic models from 10, 20 and 30.

Table 2 presents some sample topic groups and their topical words. The table shows that using seed words derived from the account owner can identify relevant contents from the list of followers.

Table 2. Sample topic groups and their topical words. ID is the topic group id

Models	IDs	Top topical words
Twitter LDA 10 topics	3	google, android, apps, mobile, galaxy, tablet
	4	samsung, galaxy, mobile, phone, android, tv, camera, smartphone
Twitter LDA 20 topics	8	galaxy, samsung, android, phone, mobile, apps, smartphone
	17	samsung, galaxy, app, tablet
	19	samsung, tv, led, mobile, smart, phone, laptop
Twitter LDA 30 topics	3	samsung, galaxy
	18	samsung, galaxy, android, google, app, phone, mobile, tablet, smartphone
	25	samsung, tv, led, camera, lcd, smart, hd

4.3 Results of SVM

The 10 fold cross-validation of the training data yields an accuracy of 88%, with class precision and recall as presented in Table 3.

Table 3. SVM 10 fold cross-validation result

	True samsungsg	True others	Class precision
Predicted samsungsg	165	13	92.7%
Predicted others	35	187	84.2%
Class recall	82.5%	93.5%	

The results of the testing data from various approaches using the SVM as compared to the baseline method – Direct Keyword Match is showed using Receiver Operator Characteristic (ROC) curves in Fig. 2. There are 3 approaches:

1. SVM_LDA: all the tweets of each follower are represented as a single feature using top topical words from Twitter LDA.
2. SVM_TF: all the tweets of each follower are represented by top frequency terms.
3. SVM_Ind: A *v* score is generated through the classification of each tweet of the follower.

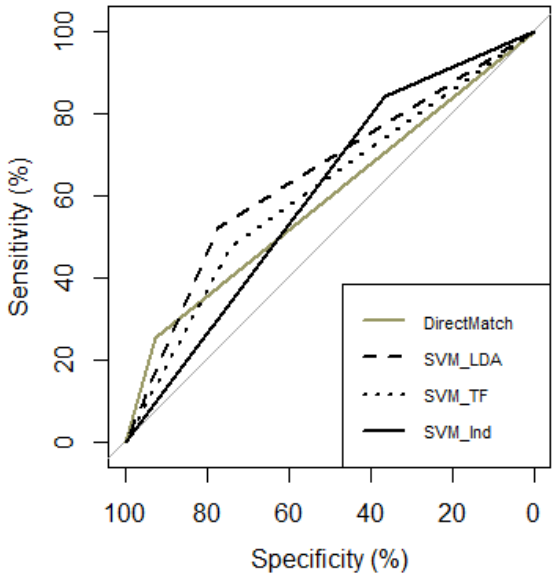

Fig. 2. ROC curves based on testing data of various approaches using the SVM.

All the three approaches performed better than the baseline Direct Keyword Match method with the third approach (SVM_Ind), which classifies individual tweets instead

of combining all the tweets in a single feature, having the higher sensitivity. This is essential as it is more capable of identifying the true high-value social audience for the account owner.

4.4 Comparison of various methods

To compare the various methods, ROC curves, as shown in Fig. 3, are plotted on all the results. It is observed that Fuzzy Keyword Match has the best result (the largest area under the curve), followed by the Twitter LDA topic modelling methods. The SVM or machine learning method has a higher sensitivity as compared to the baseline method, Direct Keyword Match, but it has not performed as well as the other methods.

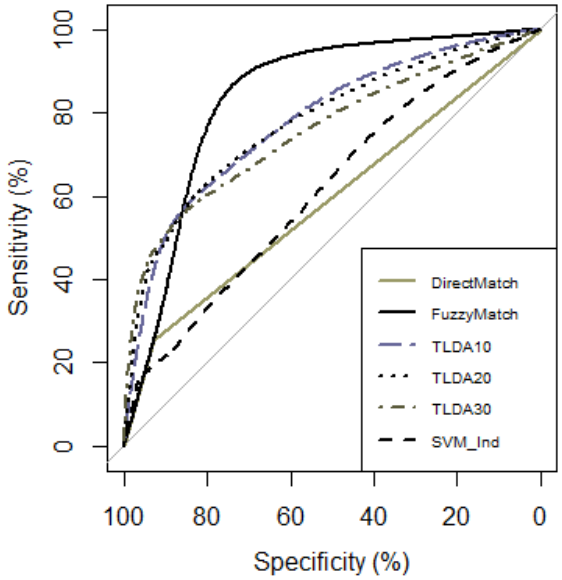

Fig. 3. ROC curves of various methods

5 Discussion

It is interesting to observe from the results that, while most account owners may think their followers are truly interested in their contents, this may not be the case as shown in the results in Table 1. It is likely that half or less than half of all the followers are tweeting similar contents to them.

One possible reason Fuzzy Keyword Match has emerged as the top performer may be due to the account chosen. “samsungsg” being a technology and mobile company,

tends to tweet contents with specific terms such as products or events. As such, it is likely that those target audience who are also interested in the similar content will be tweeting similar terms or text. For example, the *s* score (generated by the Fuzzy Keyword Match method) is the highest for both Twitter users, `follower1` and `follower2`, (as shown in Table 4) and a detailed study on their tweets indeed showed that they have tweets related to Samsung. `follower1` shared a lot of tweets on technology and mobile news, such as `"A new galaxy is born, follow @SamsungMobile for updates on the Samsung S III."`, `"Let the Smart TV experience begin | Samsung Smart TV"`. While `follower2` did share one tweet on `"3 galaxy, 2 xp, 1 iphone, 1 mac and latest 1 wins 8, under 1 roof. That wld make me? Complicated. :)"`, this user mainly tweets about daily chores. This may explain why Twitter LDA methods did not generate any high score for `follower2`. While the Fuzzy Keyword Match method seems to perform well, this may not be the case for more generic accounts such as parents groups or current affairs as the contents shared can be rather diverse and conceptual.

Even though the SVM would usually outperform most of the other methods in various other text mining studies [7], it is not the case in this study. We analyzed the top few followers with high *v* scores who are assigned by the SVM and realized that, while most of the assignments were indeed tweeting contents related to samsungsg and their scores were in-sync with scores from other methods, `follower3` wasn't. In Table 4, `follower3` was scored badly by all the other methods except for the SVM. A detailed investigation on the user's tweets only extracted one tweet – `"Having fun playing CSR Racing for Android, why not join me for FREE?"` as the rest were non-English contents. It is hence worth considering combining various methods in deriving a suitable score or index for identifying the high-value audience.

Table 4. Interesting followers identified. The highest score of each user is bolded. *s* score is generated by Fuzzy Keyword Match method, TLDA10 *t* score is generated byTwitter LDA 10 topics, TLDA20 *t* score is generated by Twitter LDA 20 topics, TLDA30 *t* score is generated by Twitter LDA 30 topics and *v* score is generated by the SVM.

Twitter name	*s* score	TLDA10 *t* score	TLDA20 *t* score	TLDA30 *t* score	*v* score
follower1	**1.0**	**1.0**	**1.0**	**1.0**	0.18
follower2	**0.9**	0.1	0.03	0.0	0.2
follower3	0.35	0.2	0.13	0.03	**1.0**

In addition, as our main intention is to find an approach to identify the high-value audience without the need to manually annotate the vast amount of tweet contents, we have used tweets from the account owner (which logically should be tweeting contents that will attract followers of similar interest) as the training data instead of using the followers' tweets. While identifying relevant tweets from the followers as the training dataset can be done through an unsupervised topic modelling method, we are

interested to explore if the content of the owner account can be used for this purpose. Analyses using followers' tweets will be studied in the future, which we expect would provide better results.

The various scores generated such as the *s* score from Fuzzy Keyword Match can be used to segment followers into groups of high-value social audience members, which a company or organization can use to engage depending on the resources available. For example, if the company or organization only has a limited amount of budget to reach out to 100 followers, the top 100 scorers would have the higher possibility of being interested than a randomly generated list of 100. In fact, a preliminary result using an average value that is built from the combination of the various scores has shown to identify 86% of the 63 high-value audience members from the 300 annotated random users. In other words, the scores derived from the various methods have high potential to be customized for segmentation of followers for social media marketing and engagement.

6 Conclusion and Future Work

In this study, we have used various text mining methods to identify the high-value social audience from a list of followers using the contents of a Twitter account owner, "samsungsg". It is assumed that those who tweeted similar contents are more likely to be interested in the owner's tweets as compared to those who have not been sharing similar contents.

Our results show that the Fuzzy Keyword Match method has produced the best performance in identifying the high-value social audience. It should be noted that achieving an accuracy of 100% for the application area of targeted marketing is unnecessary as any improvement of mass marketing is going to be beneficial for business companies.

From the result observation, it is likely that half or less of the followers are sharing similar contents as the owner, hence it makes sense to segment or identify groups of social audience who are the target audience for further engagement. Our approach in identifying this group of high-value audience members enables companies or organizations of any Twitter account owner to devise their marketing or engagement plan according to the segment or group of social audience members so as to maximize the use of allocated budgets and successfully reaching out to customers in the crowded social media space.

We have used "samsungsg" as a case study in this paper. For future work, we plan to extend it to include other account owners to verify if the observation is consistent across Twitter or if there are other features that can play a role in identifying the high-value audience. We would also like to see if the use of biologically inspired Natural Language Processing (NLP) methods [21] such as the Extreme Learning Machine (ELM) [22], which has gained increasing popularity recently, would achieve good results in unstructured text analysis. It will be of interest to explore this area and improve on the results.

References

1. Unlocking the power of social media | IAB UK, http://www.iabuk.net/blog/unlocking-the-power-of-social-media.
2. 2013 Fortune 500 - UMass Dartmouth, http://www.umassd.edu/cmr/socialmediaresearch/2013fortune500/.
3. Breslin, J.G., Passant, A., Vrandečić, D.: Social semantic web. Handbook of Semantic Web Technologies. pp. 467–506. Springer (2011).
4. Torres, D., Diaz, A., Skaf-Molli, H., Molli, P.: Semdrops: A Social Semantic Tagging Approach for Emerging Semantic Data. Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on. pp. 340–347. IEEE (2011).
5. Kondrak, G., Marcu, D., Knight, K.: Cognates can improve statistical translation models. Presented at the Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003--short papers-Volume 2 (2003).
6. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. Advances in Information Retrieval. pp. 338–349. Springer (2011).
7. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. Springer (1998).
8. Mo, J., Kiang, M.Y., Zou, P., Li, Y.: A two-stage clustering approach for multi-region segmentation. Expert Systems with Applications. 37, 7120–7131 (2010).
9. Namvar, M., Khakabimamaghani, S., Gholamian, M.R.: An approach to optimised customer segmentation and profiling using RFM, LTV, and demographic features. International Journal of Electronic Customer Relationship Management. 5, 220–235 (2011).
10. Mislove, A., Viswanath, B., Gummadi, K.P., Druschel, P.: You are who you know: inferring user profiles in online social networks. Proceedings of the third ACM international conference on Web search and data mining. pp. 251–260. ACM (2010).
11. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences. 110, 5802–5805 (2013).
12. How Ebay Uses Twitter, Smartphones and Tablets to Snap Up Shoppers, http://www.ibtimes.co.uk/how-ebay-uses-twitter-smartphones-tablets-snap-shoppers-1443441.
13. Zhang, Y., Pennacchiotti, M.: Predicting purchase behaviors from social media. Proceedings of the 22nd international conference on World Wide Web. pp. 1521–1532. International World Wide Web Conferences Steering Committee (2013).
14. Using the Twitter Search API | Twitter Developers, https://dev.twitter.com/docs/using-search.
15. Nakatani, S.: language-detection - Language Detection Library for Java - Google Project Hosting, http://code.google.com/p/language-detection/.
16. Toutanova, K., Manning, C.D.: Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. Presented at the Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13 (2000).
17. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research. 3, 993–1022 (2003).
18. Yang, M.-C., Rim, H.-C.: Identifying Interesting Twitter Contents Using Topical Analysis. Expert Systems with Applications. (2014).

19. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery. 2, 121–167 (1998).
20. Predictive Analytics, Data Mining, Self-service, Open source - RapidMiner, http://rapidminer.com/.
21. Cambria, E., Mazzocco, T., Hussain, A.: Application of multi-dimensional scaling and artificial neural networks for biologically inspired opinion mining. Biologically Inspired Cognitive Architectures. 4, 41–53 (2013).
22. Cambria, E., Huang, G.-B., Kasun, L.L.C., Zhou, H., Vong, C.-M., Lin, J., Yin, J., Cai, Z., Liu, Q., Li, K.: Extreme Learning Machines. IEEE Intelligent Systems. 28, 30–59 (2013).